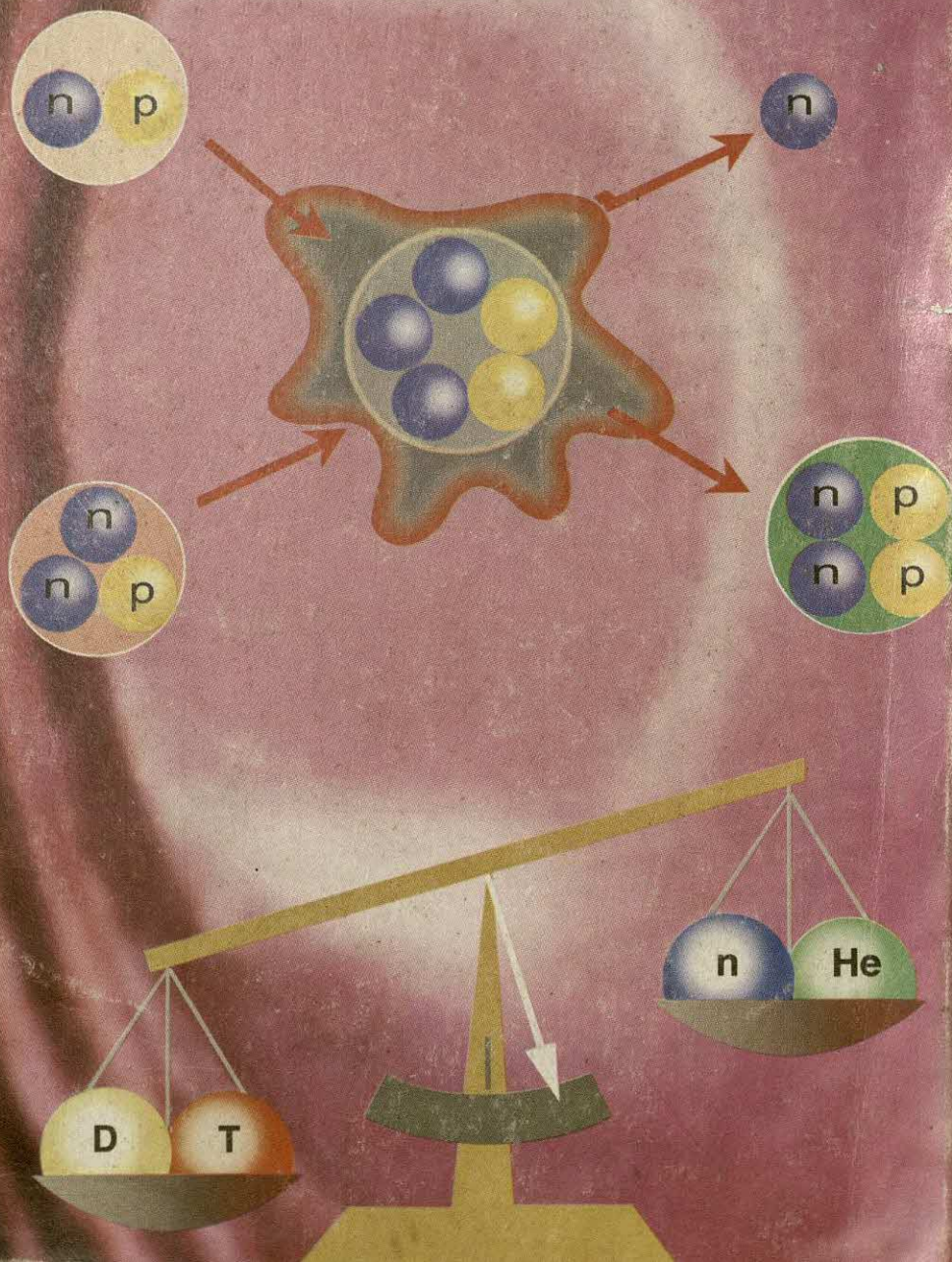


PHYSICS

TEXTBOOK FOR CLASS XII



S.C.E.R.T., W.B.
N.C.F., '2005

Text book
Code no. 018.

PHYSICS

TEXTBOOK FOR CLASS XII

AUTHORS

ARVIND KUMAR	A.S. NIGAVEKAR
B.K. SHARMA	D.P. TEWARI
P.C. JAIN	RAJARAM NITYANANDA
V.P. SRIVASTAVA	VIJAY A. SINGH
SURESH CHANDRA	

EDITORS

ARVIND KUMAR	B.K. SHARMA
P.C. JAIN	SURESH CHANDRA



राष्ट्रीय शैक्षिक अनुसंधान और प्रशिक्षण परिषद्
NATIONAL COUNCIL OF EDUCATIONAL RESEARCH AND TRAINING

First Edition

March 2003 Phalgun 1924

ISBN 81-7450-194-0

Reprinted

November 2003 Agrahayana 1925

March 2005 Phalgun 1926

November 2005 Agrahayana 1927

PD 80T VJ+13T

© National Council of Educational Research and Training, 2003

ALL RIGHTS RESERVED

- ☐ No part of this publication may be reproduced, stored in a retrieval system or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise without the prior permission of the publisher.
- ☐ This book is sold subject to the condition that it shall not, by way of trade, be lent, re-sold, hired out or otherwise disposed of without the publisher's consent, in any form of binding or cover other than that in which it is published.
- ☐ The correct price of this publication is the price printed on this page. Any revised price indicated by a rubber stamp or by a sticker or by any other means is incorrect and should be unacceptable.

OFFICES OF THE PUBLICATION DEPARTMENT, NCERT

NCERT Campus Sri Aurobindo Marg New Delhi 110 016	108, 100 Feet Road Hosdakere Halli Extension Banashankari III Stage Bangalore 560 085	Navjivan Trust Building P.O. Navjivan Ahmedabad 380 014	CWC Campus Opp. Dhankal Bus Stop Panihati Kolkata 700 114	CWC Complex Maligaon Guwahati 781 021
---	--	---	--	---

PUBLICATION TEAM

Editorial Vineet Joshi
Production Atul Saxena

Layout
Kalyan Banerjee

Rs 155.00

Printed on 70 GSM paper with NCERT watermark

Published at the Publication Department by the Secretary, National Council of Educational Research and Training, Sri Aurobindo Marg, New Delhi 110 016 and printed at Shivam Offset Pvt. Ltd., Village Asola Fateh Pur Beri, New Delhi 110 074

PUBLISHER'S NOTE

The National Council of Educational Research and Training (NCERT) has been preparing and publishing school textbooks and other educational material for children and teachers. These publications are regularly revised on the basis of feedback from students, teachers, parents, and teacher educators. Research done by the NCERT also forms the basis for updating and revision.

This book is based on the National Curriculum Framework for School Education – 2000 and the syllabi prepared in accordance with it. The Executive Committee of the NCERT, in its meeting held on 19 July 2004, discussed all aspects related to the quality of textbooks and decided that the textbooks of all subjects should undergo a quick review. In pursuance of this decision, the NCERT constituted 23 Quick Review Committees to examine all the textbooks. These committees identified various errors of conceptual, factual and linguistic nature. The review process also took note of the evaluation of textbooks undertaken earlier. The exercise has now been completed and the errors identified have been corrected. We hope that this revised edition will serve as an effective medium of teaching and learning. We look forward to your suggestions to enable us to further improve the quality of this book.

New Delhi
January 2005

SECRETARY
National Council of Educational
Research and Training

CONSTITUTION OF INDIA

Preamble

WE, THE PEOPLE OF INDIA, having solemnly resolved to constitute India into a **SOVEREIGN SOCIALIST SECULAR DEMOCRATIC REPUBLIC** and to secure to all its citizens:

JUSTICE, social, economic and political;

LIBERTY of thought, expression, belief, faith and worship;

EQUALITY of status and of opportunity; and to promote among them all

FRATERNITY assuring the dignity of the individual and the unity and integrity of the Nation;

IN OUR CONSTITUENT ASSEMBLY this twenty-sixth day of November, 1949, do **HEREBY ADOPT, ENACT AND GIVE TO OURSELVES THIS CONSTITUTION.**

PREFACE

The development of textbooks, particularly the science books, is a dynamic process in view of the changing perceptions, needs, feedback and the experiences of the students, educators and the society. The NCERT brought out the *National Curriculum Framework for School Education-2000* (NCFSE-2000) and the syllabus was accordingly revised at the school level. The new book for Class XI was written last year focusing on these changes. The present book for Class XII follows a similar pattern. This book is the result of the sincere efforts of the present team of authors with the hope that it will help in motivating and encouraging the students and teachers towards the study of the subject of Physics.

Physics is basic to the understanding of almost all the branches of science and technology. We are conscious of the fact that some of the underlying simple basic physics principles are often conceptually quite intricate. In this book, we have tried to bring in a 'conceptual coherence'. The pedagogy and the use of easily understandable language is at the core of our effort without sacrificing the **rigour** of the subject. The nature of the subject of physics is such that a certain minimum use of mathematics is a must. We have tried to develop the mathematical formulations in a logical fashion, as far as possible.

This book, as in the case of Class XI textbook, has also added some new features which, we earnestly hope, will enhance its usefulness for the students. Each chapter is provided with a **Summary** at its end for a quick overview of the contents of the chapter. This is followed by **Points to Ponder** which points out the likely misconceptions arising in the minds of students, hidden implications of certain statements/principles given in the chapter and **cautions** needed in applying the knowledge gained from the chapter. Students will find it exciting to think and apply their mind on these **points**. Further, a large number of **solved examples** are included in the text in order to clarify the concepts and/or to illustrate the application of these concepts in everyday real-life situations. Occasionally, historical perspective has been included to share the excitement of sequential development of the subject of physics. Some **Boxed** items are introduced in many chapters either for this purpose or to highlight some special features of the contents requiring additional attention of the learners.

Special attention has been paid for providing illustrative figures. To increase the clarity, many figures are drawn in two 'colours'. A large number of Exercises are given in the end of each chapter. Some of these are from real-life situations. Answers and 'hints' to solve some of these are also included. Students are urged to solve and in doing so, they may find these Exercises very educative. In the entire book, SI units have been used.

Completing this book has only been possible because of the spontaneous and continuous support of many people. We express our gratitude to the Director, NCERT for entrusting us with the task of preparing this textbook as a part of the national effort for improving science education. He encouraged us by providing all the administrative and academic inputs. The Head, Department of Education in Science and Mathematics, NCERT willingly helped us in our endeavour in every possible way. It was a pleasure to work with him. The coordinator, apart from being an author, ably co-ordinated the entire task and

activities of the Writing team, besides the overall supervision of publication of the final manuscript.

The present text got excellent suggestions from the teachers and experts for the improvement during the Review Workshop organised to discuss and refine the first draft. Occasionally, some portions from the old NCERT book, particularly appreciated by students/teachers, have been adopted/adapted and retained in the present book for the benefit of coming generation of learners.

We welcome suggestions and comments from our users, especially the students and teachers. We wish our young readers a happy journey to the exciting realm of physics.

SURESH CHANDRA
Chairman
Writing team

ACKNOWLEDGEMENTS

The National Council of Educational Research and Training is grateful to the members of the Writing Team for their valuable contributions in the development of manuscript of this textbook. The Council expresses its gratefulness to Professor Suresh Chandra (*Chairman*), Emeritus Scientist, Banaras Hindu University, Varanasi; Shri Rajaram Nityananda, *Director*, National Centre for Radio-Astrophysics, Pune; Professor P.C. Jain, University of Delhi; Professor Vijay A. Singh, IIT, Kanpur; Professor Arvind Kumar, *Director*, Homi Bhabha Centre for Science Education, Mumbai; Professor A.S. Nigavekar, *Chairman*, UGC, New Delhi; Professor D.P. Tewari, (*Retd*), IIT, New Delhi; and Dr V.P. Srivastava, *Reader*; Professor B.K. Sharma (*Coordinator*), DESM, NCERT, New Delhi.

The Council also acknowledges the valuable contributions of the following participants of the Review Workshop in the finalisation of this book: Professor S.S. Kushwaha, (*Retd*), Banaras Hindu University, Varanasi; Shri P.C. Agarwal, *Reader*, RIE, Bhubaneswar; Smt. Rachna Garg, *Lecturer*, RIE, Bhopal; Shri Ram Prakash Gupta, *PGT*, Rajkiya Pratibha Vikas Vidyalaya, New Delhi; Shri S.C. Garg, *Lecturer*, Government P.G. College, Jind, Haryana; Shri Mukesh Kumar Gandhi, *PGT*, Delhi Public School, Ghaziabad; Shri Suraj Prakash, *Principal*, CRPF Public School, Delhi; Shri Vivek Misra, *PGT*, Mount Carmel School, New Delhi; Shri Sher Singh, *PGT*, NDMC Navyug School, Lodhi Road, New Delhi; Shri A.K. Das, *PGT*, St. Xavier's Senior Secondary School, Delhi; Shri B.M. Mudgal, *PGT*, Air Force Golden Jubilee Institute, Delhi; Shri Anil Kumar, *Principal*, Government Boys Senior Secondary School, Khera Khurd, Delhi; Shri S.N. Prabhakara, *Headmaster*, Demonstration Multipurpose School, RIE, Ajmer; Smt. Yashu Kumar, *PGT*, Kulachi Hansraj Model School, Delhi; Shri V.K. Gautum, *Principal*, Kendriya Vidyalaya, Gole Market, New Delhi; Shri Sanjay Yadav, *PGT*, The Mother's International School, New Delhi; Shri Suresh Kumar, *PGT*, Delhi Public School, Dwarka, New Delhi; Shri R.S. Dass, *Vice-Principal*, Balwant Ray Mehta Vidya Bhawan Senior Secondary School, New Delhi; Shri Rajesh Kumar, *Lecturer*, SCERT, DIET, Pitampura, Delhi; Smt. Anuradha Mathur, *PGT*, Modern School, New Delhi; Professor K.C. Sharma, Himachal Pradesh University, Shimla; Professor H.C. Pradhan, *Dean*, Homi Bhabha Centre for Science Education, Mumbai; Shri J.P. Agarwal, *Principal*, (*Retd*), 3, Shakti Apartment, New Delhi; Shri M.M. Sahajwani, *PGT*, Demonstration Multipurpose School, RIE, Mysore; Professor R.M.P. Jaiswal, (*Retd*), Urban Estate, Kurukshetra; Professor Vinod Prakash, University of Allahabad; Shri R.K. Prasad, *PGT*, Jawahar Navodaya Vidyalaya, Delhi; and Shri Gagan Gupta, *Reader*, DESM, NCERT, New Delhi.

The Council also gratefully acknowledges the feedback and suggestions of the following members of the Quick Review Committee for the improvement of this book: Professor Deepak Kumar, (*Chairperson*), School of Physical Sciences, Jawaharlal Nehru University, New Delhi; Dr Rabinder Nath Kakarya, *PGT*, Darbari Lal DAVMS, Pitampura, New Delhi; Shri A.K. Das, *PGT*, St. Xavier's Senior Secondary School, Delhi; Smt. Yashu Kumar, *PGT*, Kulachi Hansraj Model School, Delhi; Shri Danesh Gupta, *PGT*, Navyug School, Sarojini Nagar,

New Delhi; Smt. Anuradha Mathur, PGT, Modern School, New Delhi; Shri Suresh Kumar, PGT, Delhi Public School, Dwarka, New Delhi; Smt. Sharmila Banerjee, PGT, The Mother's International School, New Delhi; Smt. Chitra Goel, PGT, Rajkiya Pratibha Vikas Vidyalaya, Tyagraj Nagar, New Delhi; Smt. Manjusha Rawat, PGT, Kendriya Vidyalaya, Andrews Ganj, New Delhi; Smt. Neelam Sehgal, PGT, Kendriya Vidyalaya, JNU Campus, New Delhi; Dr Rupamanjari Ghosh, School of Physical Sciences, Jawaharlal Nehru University, New Delhi; Professor Pankaj Sharan, Jamia Millia Islamia, New Delhi; Professor H.C. Jain, Head, DESM, RIE, Ajmer; Smt. Ramneek Kapoor, PGT, Jaspal Public School, New Delhi; Professor R.S. Gupta, South Campus, University of Delhi; Shri Surendra Singh, PGT, Kendriya Vidyalaya, Janakpuri, New Delhi and Dr V.P. Srivastava, Reader; Dr S.K. Dash, Reader; Professor B.K. Sharma, (Coordinator), DESM, NCERT, New Delhi.

CONSTITUTION OF INDIA

Part IV A (Article 51 A)

Fundamental Duties

Fundamental Duties – It shall be the duty of every citizen of India —

- (a) to abide by the Constitution and respect its ideals and institutions, the National Flag and the National Anthem;
- (b) to cherish and follow the noble ideals which inspired our national struggle for freedom;
- (c) to uphold and protect the sovereignty, unity and integrity of India;
- (d) to defend the country and render national service when called upon to do so;
- (e) to promote harmony and the spirit of common brotherhood amongst all the people of India transcending religious, linguistic and regional or sectional diversities; to renounce practices derogatory to the dignity of women;
- (f) to value and preserve the rich heritage of our composite culture;
- (g) to protect and improve the natural environment including forests, lakes, rivers, wildlife and to have compassion for living creatures;
- (h) to develop the scientific temper, humanism and the spirit of inquiry and reform;
- (i) to safeguard public property and to abjure violence;
- (j) to strive towards excellence in all spheres of individual and collective activity so that the nation constantly rises to higher levels of endeavour and achievement;
- (k) who is a parent or guardian, to provide opportunities for education to his child or, as the case may be, ward between the age of six and fourteen years.

CONSTITUTION OF INDIA

Part III (Articles 12 – 35)

(Subject to certain conditions, some exceptions
and reasonable restrictions)

guarantees these

Fundamental Rights

Right to Equality

- before law and equal protection of laws;
- irrespective of religion, race, caste, sex or place of birth;
- of opportunity in public employment;
- by abolition of untouchability and titles.

Right to Freedom

- of expression, assembly, association, movement, residence and profession;
- of certain protections in respect of conviction for offences;
- of protection of life and personal liberty;
- of free and compulsory education for children between the age of six and fourteen years;
- of protection against arrest and detention in certain cases.

Right against Exploitation

- for prohibition of traffic in human beings and forced labour;
- for prohibition of employment of children in hazardous jobs.

Right to Freedom of Religion

- freedom of conscience and free profession, practice and propagation of religion;
- freedom to manage religious affairs;
- freedom as to payment of taxes for promotion of any particular religion;
- freedom as to attendance at religious instruction or religious worship in educational institutions wholly maintained by the State.

Cultural and Educational Rights

- for protection of interests of minorities to conserve their language, script and culture;
- for minorities to establish and administer educational institutions of their choice.

Right to Constitutional Remedies

- by issuance of directions or orders or writs by the Supreme Court and High Courts for enforcement of these Fundamental Rights.



CONTENTS

PUBLISHER'S NOTE

PREFACE

CHAPTER 1

ELECTRIC CHARGES AND FIELDS

1.1	Introduction	1
1.2	Electric Charge	3
1.3	Electricity and Matter	4
1.4	Charging by Induction	4
1.5	Coulomb's Law	5
1.6	Basic properties of Electric Charge	7
1.7	Multiple charges: The Superposition Principle	9
1.8	Electric Field	10
1.9	Electric Dipole	13
1.10	Dipole in a Uniform External Field	16
1.11	Electric Field Lines	17
1.12	Electric Flux	19
1.13	Gauss's Theorem	20
1.14	Continuous Charge Distribution	23
1.15	Applications of Gauss's Theorem	24

CHAPTER 2

ELECTROSTATIC POTENTIAL AND CAPACITANCE

2.1	Introduction	36
2.2	Electrostatic Potential	39
2.3	Potential due to a Point Charge	39
2.4	Potential due to an Electric Dipole	40
2.5	Potential due to a System of Charges	41
2.6	Equipotential Surfaces	42
2.7	Potential Energy of a System of Charges	43
2.8	Potential Energy in an External Field	44
2.9	Electrostatics of Conductors	47
2.10	Capacitors and Capacitance	49
2.11	The Parallel Plate Capacitor	50
2.12	Combinations of Capacitors	51
2.13	Energy Stored in a Capacitor	53
2.14	Dielectrics and Polarisation	54
2.15	Effect of Dielectric on Capacitance	56
2.16	Van de Graaff Generator	58

CHAPTER 3**CURRENT ELECTRICITY**

3.1	Introduction	68
3.2	Electric Current	70
3.3	Electromotive Force (emf) and Voltage	70
3.4	Resistance and Resistivity	71
3.5	Origin of Resistivity	74
3.6	Temperature Dependence of Resistivity	77
3.7	Limitations of Ohm's Law	79
3.8	Superconductivity	81
3.9	Resistors in Series and in Parallel	81
3.10	Electric Circuits and Kirchhoff's rules	83
3.11	Measurement of Voltages, Currents, and Resistances	87

CHAPTER 4**THERMAL AND CHEMICAL EFFECTS OF CURRENTS**

4.1	Introduction	105
4.2	Heating Effects: Joule's Law	107
4.3	Practical Applications of Joule's Heating	107
4.4	Chemical Effects of Current	109
4.5	Electrochemical Cells	114
4.6	Thermoelectricity	119
4.7	Applications of Thermoelectricity	123

CHAPTER 5**MOVING CHARGES AND MAGNETISM**

5.1	Introduction	130
5.2	The Law of Biot and Savart	133
5.3	Evaluation of the Magnetic Field	135
5.4	Ampere's Circuital Law	138
5.5	The Solenoid and the Toroid	140
5.6	The Lorentz Force	143
5.7	The Cyclotron	145
5.8	The Ampere	147
5.9	The Current Loop as a Magnetic Dipole	148
5.10	The Moving Coil Galvanometer (MCG)	153

CHAPTER 6**MAGNETISM AND MATTER**

6.1	Introduction	163
6.2	The Bar Magnet	166
6.3	Magnetism and Gauss's Law	170
6.4	The Earth Magnetism	170
6.5	Magnetisation and Magnetic Intensity	173

6.6	Magnetic Properties of Materials	174
6.7	Permanent Magnets and Electromagnets	178

CHAPTER 7

ELECTROMAGNETIC INDUCTION

7.1	Introduction	188
7.2	The Experiments of Faraday and Henry	190
7.3	Faraday's Laws of Induction	191
7.4	Lenz's Law	192
7.5	Motional emf and Faraday's Law	192
7.6	Energy Consideration: A Quantitative study	194
7.7	Eddy Currents	196
7.8	Inductance	197
7.9	AC Generator	200

CHAPTER 8

ALTERNATING CURRENT

8.1	Introduction	211
8.2	AC Voltage applied to a Resistor	213
8.3	Representation of AC Current and Voltage by Rotating Vectors-Phasors	214
8.4	AC Voltage applied to an Inductor	215
8.5	AC Voltage applied to a Capacitor	217
8.6	AC Voltage applied to a Series LCR Circuit	218
8.7	Power in AC Circuits: The Power Factor	223
8.8	LC Oscillations	225
8.9	Transformers	227

CHAPTER 9

ELECTROMAGNETIC WAVES

9.1	Introduction	236
9.2	Electromagnetic Waves	238
9.3	Electromagnetic Spectrum	243
9.4	Propagation of Electromagnetic Waves in Atmosphere	245

CHAPTER 10

RAY OPTICS AND OPTICAL INSTRUMENTS

10.1	Introduction	251
10.2	Reflection of Light by Spherical Mirrors	255
10.3	Refraction	258
10.4	Total Internal Reflection	259
10.5	Refraction at Spherical Surfaces and by Lenses	261
10.6	Refraction in a Prism	266
10.7	Dispersion by a Prism	267

10.8 Spectrometer	268
10.9 Light in Nature	269
10.10 Optical Instruments	270

CHAPTER 11

WAVE OPTICS

11.1 Introduction	289
11.2 Wave fronts, Rays and Huygens' Principle	292
11.3 Coherent and Incoherent addition of Light Waves	296
11.4 Interference	299
11.5 Diffraction	303
11.6 Polarisation	308

CHAPTER 12

DUAL NATURE OF RADIATION AND MATTER

12.1 Introduction	319
12.2 Electron Emission	321
12.3 Photoelectric Effect	321
12.4 Experimental Study of Photoelectric Effect	321
12.5 Photoelectric Effect and Wave Theory of Light	324
12.6 Einstein's Photoelectric Equation: Energy Quantum of Radiation	325
12.7 The Photon	326
12.8 Photo-Cell	327
12.9 The Wave Nature of Matter	328
12.10 The Davisson and Germer Experiment	331
12.11 The Electron Microscope	332

CHAPTER 13

ATOMS

13.1 Introduction	343
13.2 Alpha-Particle Scattering and Rutherford's Model of Atom	346
13.3 Atomic Spectra	348
13.4 Energy Quantisation	349
13.5 Bohr's Atomic Model and the Hydrogen Spectrum	351
13.6 Emission and Absorption Spectrum	356
13.7 Many Electron Atoms	356
13.8 X-Rays and the Atomic Number	359
13.9 Spontaneous and Stimulated Emission - Maser and Laser	363

CHAPTER 14

NUCLEI

14.1 Introduction	379
14.2 Atomic masses: Composition of Nucleus	380
14.3 Size of the Nucleus	381

14.4	Nuclear Binding Energies	382
14.5	Nuclear Energy Levels	383
14.6	The Nuclear Force	383
14.7	Nuclear Stability - Radioactivity	384
14.8	Nuclear Reactions	391

CHAPTER 15

SEMICONDUCTOR ELECTRONICS: MATERIALS, DEVICES AND SIMPLE CIRCUITS

15.1	Introduction	409
15.2	Semiconductor Physics: Some Basics	411
15.3	Electrical Conduction in Semiconductors	411
15.4	Energy Band description of Solids-Metals, Insulators and Semiconductors	418
15.5	p-n Junction - Basic Unit of all Semiconductor Devices	422
15.6	Voltage-Current (V-I) characteristics of a p-n Junction Diode	426
15.7	Application of p-n Diode as a Rectifier	428
15.8	Special purpose p-n Junction Diodes	429
15.9	Transistors	435
15.10	Digital Electronics and Logic Gates	441
15.11	Integrated Circuits	446

CHAPTER 16

COMMUNICATION SYSTEMS

16.1	Introduction	453
16.2	Types of Communication Systems	455
16.3	Modulation: An Important Step of Communication Systems	455
16.4	Digital Communication and Quantisation of Message Signal	460
16.5	Data and Document Transmission: Fax and Modem	461
16.6	Communication Channels	462
16.7	Space Communication	462
16.8	Satellite Communication	465
16.9	Remote Sensing: An Application of Satellite Communication	466
16.10	Line Communication	467
16.11	Optical Communication	469
16.12	Optical Fibre	471

APPENDICES	477
-------------------	-----

ANSWERS	479
----------------	-----

BIBLIOGRAPHY	548
---------------------	-----

COVER DESIGN

Anand D. Ghaisas,

**Homi Bhabha Centre for Science Education
(Tata Institute of Fundamental Research), Mumbai.**

(Adapted from the website www.plasmas.org)

Deuterium and tritium fuse at high temperatures to produce helium and a neutron. The excess mass of the reactants over the mass of the products appears as energy. In a fusion test reactor based on the technique of magnetic confinement (Tokamak), a hot dense plasma is confined for sufficient time by suitably designed magnetic fields to enable the nuclei to fuse and generate energy. The background to the schematic reaction shown is a view of the plasmas during operation.

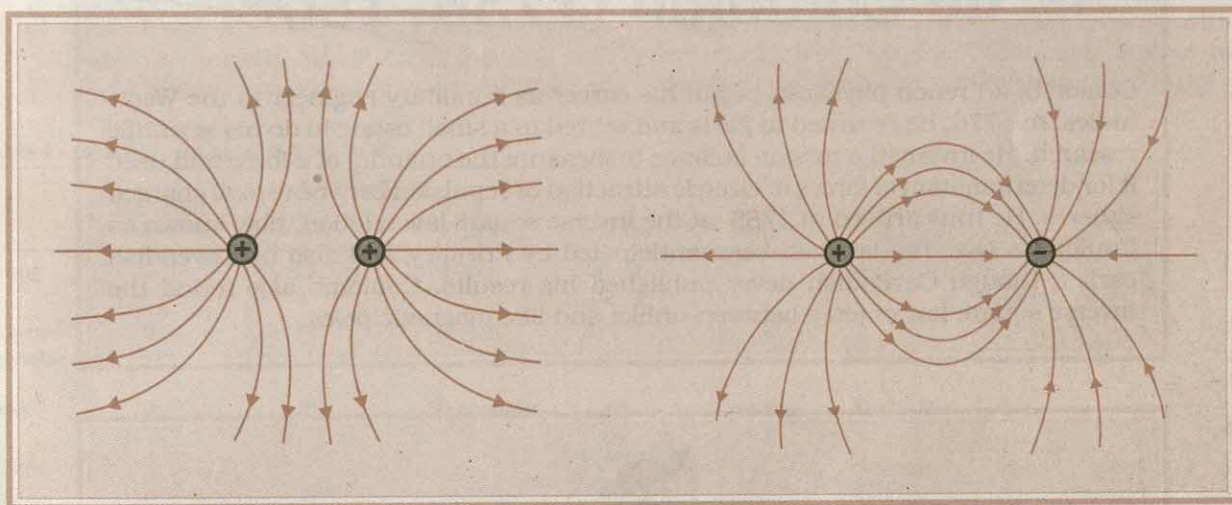
BACK COVER

**National Centre for Radio Astrophysics
(Tata Institute of Fundamental Research), Pune**

A parabolic reflecting antenna for receiving cosmic radio waves of wavelengths between 0.2 and 2 metres, located in Junnar Taluk, Pune rural district, Maharashtra. This is one of the thirty such antennas, each 45 metres in diameter, which make up the Giant Metrewave Radio Telescope (GMRT) built and operated by the National Centre for Radio Astrophysics (NCRA) of the Tata Institute of Fundamental Research (TIFR), India.

CHAPTER ONE

ELECTRIC CHARGES AND FIELDS



1.1 INTRODUCTION

The science of electricity and magnetism is basic to the modern technological civilisation. Electric power, telecommunication, radio and television, and so many of the practical appliances that we use in daily life are based on the principles of this science. Yet only two hundred years ago, electric and magnetic phenomena seemed unusual and curious, unrelated to the ordinary phenomena that we see around us. However, our understanding of nature has vastly improved in the last two centuries. We now know that the electric force is as pervasive as the gravitational force and that these two fundamental forces underlie nearly all natural phenomena (Chapter 1, Class XI). This is so because any piece of matter, even if electrically neutral as a whole, actually consists of elementary charged constituents. It is, therefore, not surprising that any microscopic understanding of the properties of matter requires the study of electric and magnetic phenomena.

Electric and magnetic phenomena are generally bracketed together, since both derive from charged particles. Magnetism, as you will learn in Chapter 5, arises from charges in motion. Charged particles in motion exert both electric and magnetic types of forces on each other. Because of this inseparable nature, electricity and magnetism are regarded as two aspects of the general subject called *electromagnetism*. However, in the frame of reference where all charges are at rest, the forces are purely electrical. The subject of *Electrostatics*, as the name suggests, deals with the physics of charges at rest. *Electrostatics* forms the content of this and the next chapter.



Charles Augustin de Coulomb (1736–1806)

Coulomb, a French physicist, began his career as a military engineer in the West-Indies. In 1776, he returned to Paris and retired to a small estate to do his scientific research. He invented a torsion balance to measure the quantity of a force and used it for determination of forces of electric attraction or repulsion between small charged spheres. He thus arrived in 1785 at the inverse square law relation, now known as Coulomb's law. The law had been anticipated by Priestley and also by Cavendish earlier, though Cavendish never published his results. Coulomb also found the inverse square law of force between unlike and like magnetic poles.



Henry Cavendish (1731-1810)

Outstanding physicist and chemist. He became wealthy by inheritance at age 40, but remained a recluse all his life. His results on electricity were unpublished and remained unknown till Maxwell edited and published them around 1875. Apart from the law of electrostatic attraction and repulsion, he invented the torsion balance to measure the Newtonian gravitational constant by a laboratory experiment, and thus weighed the earth! He also did experiments on the composition of air, and properties of hydrogen. The Cavendish laboratory of Cambridge University was set up in his honour in 1871.

1.2 ELECTRIC CHARGE

We have used the term 'electric charge' in section 1.1, since you already have some elementary ideas about charge. Charge comes in two varieties. Like charges repel and unlike charges attract each other.

These facts follow from simple experiments in frictional electricity (charging by friction) :

- Take a glass rod, rub its one end with a silk cloth and suspend it by a long thread. Take another glass rod, rub its end with a silk cloth and bring the end close to the end of the first rod. The suspended rod will swing away, showing repulsion [Fig. 1.1(a)].
- Repeat the experiment, this time with two plastic rods each rubbed with fur. You will again notice repulsion between the ends of the two rods [Fig. 1.1(b)].
- Repeat the experiment with one modification. Take one of the rods to be a glass rod rubbed with silk and the other a plastic rod rubbed with fur. In this case, the rods will show attraction [Fig. 1.1(c)].

The simple interpretation of these observations is that a glass rod rubbed with silk acquires one kind of something that we now call **charge**. The plastic rod rubbed with fur acquires another kind of charge. Like charges repel and unlike charges attract each other.

These experiments can be done on other pairs of bodies charged as before by friction. In each case you can interpret the observations by labelling the charge as either the kind acquired by a glass rod when rubbed with silk cloth or the kind acquired by a plastic rod when rubbed

with fur. This shows that there is no third variety of charge. By convention, the charge on the glass rod when rubbed with silk is termed **positive**, while that on the plastic rod when rubbed with fur is termed **negative**.

There is another related observation. When a plastic rod rubbed with fur acquires negative charge, the fur itself is found experimentally to acquire a positive charge. Moreover, if the rod and fur are brought into contact with one another, the charge is found to disappear from both. They no longer attract or repel other charged objects. Similar observations are true for the glass rod rubbed with silk cloth. The simple reason for these observations is that neutral bodies have equal amounts of positive and negative charges. When two bodies are rubbed against each other, some charge is transferred from one body to the other (usually, as we shall learn later, it is the negative charge of the more mobile electrons that flows). If by friction, negative charge is transferred, say, from A to B, A acquires a positive charge and B acquires a negative charge. On bringing back A and B in contact, the excess negative charge on B flows back to A and the two bodies become neutral. This suggests that charge is neither created nor destroyed but can be transferred from one body to another.

A simple apparatus to detect charge on a body is the leaf electroscope (Fig. 1.2). It consists of a vertical metal rod housed in a box, with two thin gold leaves attached to its bottom end. When a charged object touches the metal knob at the top of the rod, charge flows on to the leaves and they diverge. The degree of divergence is an indicator of the amount of charge.

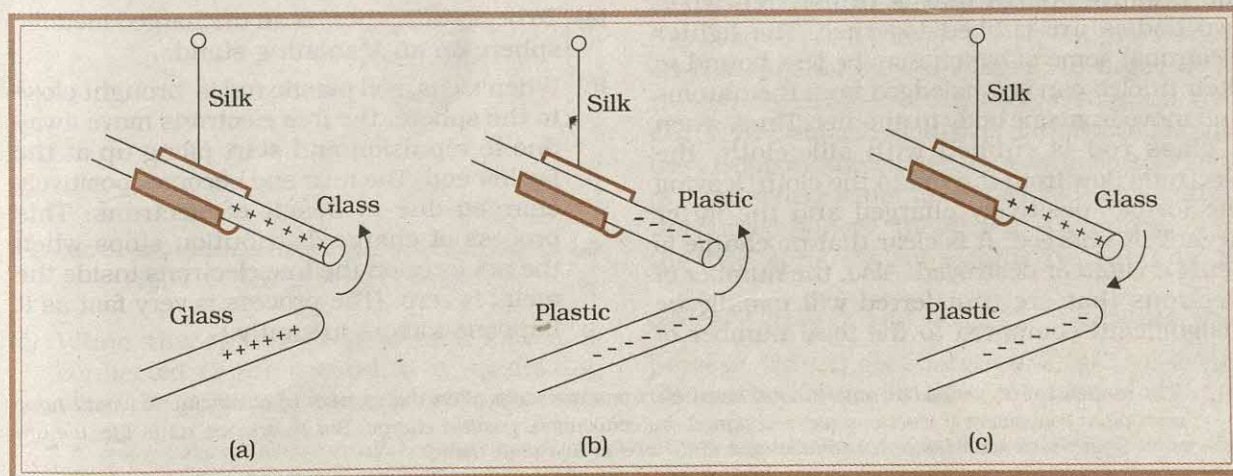


Fig. 1.1 Like charges repel and unlike charges attract each other.

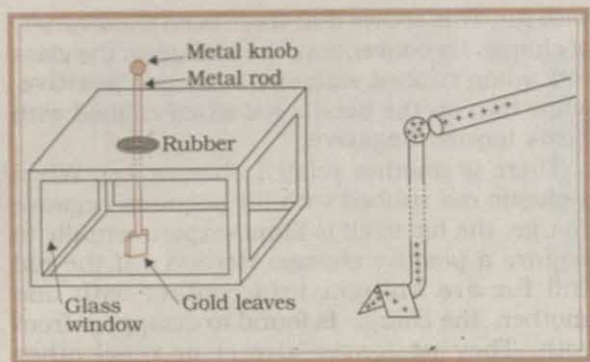


Fig. 1.2 The leaf electroscope.

1.3 ELECTRICITY AND MATTER

It is useful to have in mind the atomic picture of matter that you will learn in more detail later. The basic unit of all matter is an atom. Each atom has a small core (nucleus) which accounts for most of its mass, and lighter particles (electrons) orbiting around the nucleus. The nucleus and the electrons have opposite electric charges. With the convention already mentioned in section 1.2, the nucleus has positive charge and electrons are negatively charged*. The nucleus is made of positively charged protons and neutral particles called neutrons. The electron is much lighter than a proton or neutron; its mass is approximately two thousandth of the mass of a proton or neutron. The electric charge of a proton is exactly equal in magnitude to the charge of an electron. Thus, in a neutral atom (or neutral matter), the total number of protons equals the total number of electrons.

Now you can understand why it is usually the negative charge that is transferred when two bodies are rubbed together. The lighter electrons, some of which may be less bound to their nuclei, can get dislodged from their atoms and move from one body to another. Thus, when a glass rod is rubbed with silk cloth, the electrons flow from the rod to the cloth, leaving the former positively charged and the latter negatively charged. It is clear that no charge is being created or destroyed. Also, the number of electrons that are transferred will usually be insignificant compared to the total number of

electrons or protons in the body. Thus, the charge acquired by friction is a very small fraction of the total positive and negative charge content in a body.

Some substances readily allow passage of electricity through them, others do not. For example, if you connect a charged plastic rod to an uncharged pith ball by means of a copper wire, the ball almost instantly gets charged. This is because part of the charge of the rod goes through the wire to the ball. This will not happen if the copper wire is replaced by a nylon wire. A copper wire is a conductor of electricity while a nylon wire is an insulator. Most substances fall into one of the two classes : conductors and insulators**. In conductors, some of the outer electrons of individual atoms get detached from them and move almost freely inside the substance. Insulators do not have (or have negligible number of) free electrons. It is because of these free electrons that a conductor readily allows passage of charge.

1.4 CHARGING BY INDUCTION

We have seen that a body may be charged by putting it in contact with another charged body either directly or by means of a conductor. Thus, when a charged plastic rod is in contact with a pith ball or connected to it by a copper wire, it transfers some of its charge to the ball. In charging by induction, a charged body A imparts to another body B a charge of opposite sign without any kind of contact between A and B. Since there is no contact, the body A does not lose any charge.

Figure 1.3 shows the steps involved in charging a metallic sphere by induction:

- To begin with, there is an uncharged metallic sphere on an insulating stand.
- When a charged plastic rod is brought close to the sphere, the free electrons move away due to repulsion and start piling up at the farther end. The near end becomes positively charged due to deficit of electrons. This process of charge distribution stops when the net force on the free electrons inside the metal is zero. (The process is very fast as it happens almost instantly.)

* This convention is somewhat unfortunate, since electrons are more often the carriers of electricity. It would have been more convenient if electrons were assigned, by convention, positive charge. But in science, as in life, we are sometimes stuck with historical conventions and have to live with them!

** There is a third important category – the semiconductors, that we do not discuss here. Semiconductors are described in Chapter 15.

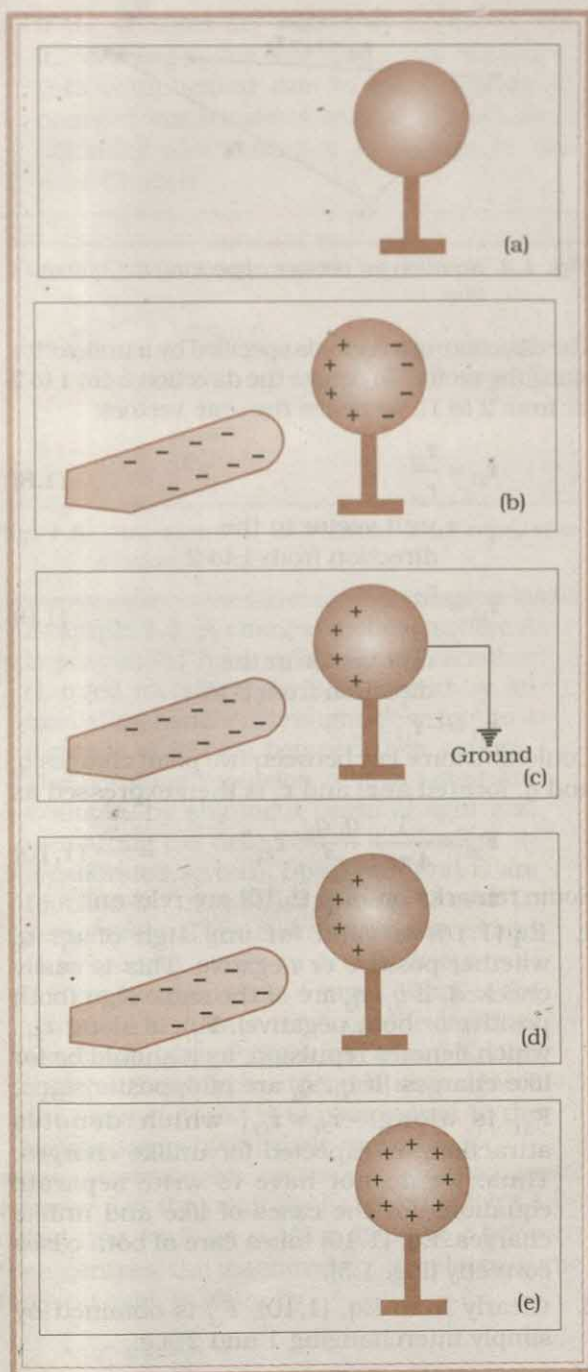


Fig. 1.3 Charging by induction.

- (c) When the sphere is grounded i.e., it is connected to the ground by a conducting wire, the negative charge (electrons) flows

to the ground. The positive charge at the near end remains held there due to the attractive force of the external charge.

- (d) When the sphere is disconnected from the ground, the positive charge continues to be held at the near end.
 (e) When the plastic rod is removed, the positive charge spreads uniformly over the sphere.

Similar steps are involved if a positively charged rod is used for charging the metallic sphere by induction. Note, however, that in this case the electrons flow from the ground to the sphere in step (c).

1.5 COULOMB'S LAW

Coulomb's law is a quantitative statement on the force between two **point charges**. When the linear sizes of charged bodies are much smaller than the distance separating them, the size may be ignored and the charged bodies are called **point charges**. Coulomb (1736 - 1806) measured the force between two **point charges** and found that it varied inversely as the square of the distance between the charges and was directly proportional to the product of the magnitude of the two charges. Thus, if two point charges q_1 , q_2 are separated by a distance r , the magnitude of the force (F) between them is given by

$$F = k \frac{|q_1 q_2|}{r^2} \quad (1.1)$$

How did Coulomb arrive at this law from his experiment? Coulomb used a torsion balance* for measuring the force between two charged metallic spheres. When the separation between two spheres is much larger than the radius of each sphere, the charged spheres may be regarded as point charges. However, the charges on the spheres were unknown, to begin with. How then could he discover a relation like Eq. (1.1)? Coulomb thought of the following simple way: Suppose the charge on a metallic sphere is q . If the sphere is put in contact with an identical uncharged sphere, the charge will spread over the two spheres. By symmetry, the charge

on each sphere will be $\frac{q}{2}$ **. Repeating this process, we can get charges $q/2$, $q/3$, $q/4$, etc. Comparing forces for different pairs of charges

* A torsion balance, whose details we omit, is a sensitive device to measure force. It was also used later by Cavendish to measure the very feeble gravitational force between two objects, to verify Newton's Law of Gravitation.

** Implicit in this is the assumption of additivity of charges: two charges ($q/2$ each) add up to make a total charge q .

at different distances, Coulomb arrived at the relation, Eq. (1.1).

Note the wonderful thing. Coulomb discovered his law without knowing the explicit magnitude of the charge. In fact, it is the other way round. Coulomb's law can be employed to furnish a definition for a unit of charge. In the relation, Eq. (1.1), k is so far arbitrary. We can choose any value of k that we like. The choice of k determines the size of the unit of charge. In SI units, the value of k is about 9×10^9 . The unit that results from this choice is called Coulomb (C). Putting this value of k in Eq. (1.1), we see that for $q_1 = q_2 = 1 \text{ C}$, $r = 1 \text{ m}$,

$$F = 9 \times 10^9 \text{ N}.$$

That is, 1 C is the charge which when placed at a distance of 1 m from another charge of the same magnitude in vacuum experiences an electrical force of repulsion of magnitude $9 \times 10^9 \text{ N}$ *. 1C is evidently too big a unit. In practice, in electrostatics, one uses smaller units like $1 \text{ mC} = (10^{-3} \text{ C})$ or $1 \mu\text{C} = (10^{-6} \text{ C})$. The constant k in Eq. (1.1) is usually put as

$$k = \frac{1}{4\pi\epsilon_0} \quad (1.2)$$

for later convenience, so that Coulomb's law is written as

$$F = \frac{1}{4\pi\epsilon_0} \frac{|q_1 q_2|}{r^2} \quad (1.3)$$

In terms of ϵ_0 , some of the later equations are simpler. The value of ϵ_0 in SI units is

$$\epsilon_0 = 8.854 \times 10^{-12} \text{ C}^2 \text{ N}^{-1} \text{ m}^{-2}$$

We can express Coulomb's law better as a vector relation. Let us first establish our notation :

$$\begin{aligned} \text{position vector of charge } q_1 &= \mathbf{r}_1 \\ \text{position vector of charge } q_2 &= \mathbf{r}_2 \\ \text{force on } q_1 \text{ by } q_2 &= \mathbf{F}_{12} \\ \text{force on } q_2 \text{ by } q_1 &= \mathbf{F}_{21} \end{aligned} \quad (1.4)$$

The two point charges q_1 and q_2 have been numbered 1 and 2 for convenience. Further, denote the vector leading from 1 to 2 by \mathbf{r}_{21} :

$$\mathbf{r}_{21} = \mathbf{r}_2 - \mathbf{r}_1 \quad (1.5)$$

In the same way, the vector leading from 2 to 1 is denoted by \mathbf{r}_{12} :

$$\mathbf{r}_{12} = \mathbf{r}_1 - \mathbf{r}_2 = -\mathbf{r}_{21} \quad (1.6)$$

The magnitude of the vectors \mathbf{r}_{21} and \mathbf{r}_{12} is denoted by r_{12} :

$$|\mathbf{r}_{21}| = |\mathbf{r}_{12}| = r_{12} = r_{21} \quad (1.7)$$

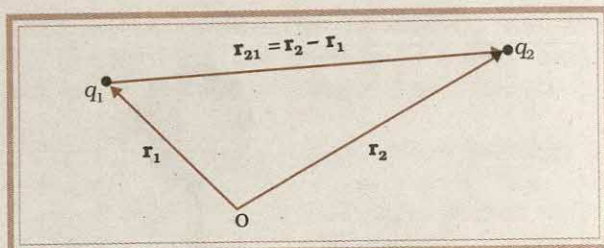


Fig. 1.4 Notation for vectors appearing in Coulomb's law.

The direction of a vector is specified by a unit vector along the vector. To denote the direction from 1 to 2 (or from 2 to 1), we define the unit vectors:

$$\begin{aligned} \hat{\mathbf{r}}_{21} &= \frac{\mathbf{r}_{21}}{r_{21}} \\ &= \text{unit vector in the} \\ &\quad \text{direction from 1 to 2} \end{aligned} \quad (1.8)$$

$$\begin{aligned} \hat{\mathbf{r}}_{12} &= \frac{\mathbf{r}_{12}}{r_{12}} \\ &= \text{unit vector in the} \\ &\quad \text{direction from 2 to 1} \end{aligned} \quad (1.9)$$

$$\hat{\mathbf{r}}_{21} = -\hat{\mathbf{r}}_{12}$$

Coulomb's force law between two point charges q_1 and q_2 located at \mathbf{r}_1 and \mathbf{r}_2 is then expressed as

$$\mathbf{F}_{21} = \frac{1}{4\pi\epsilon_0} \frac{q_1 q_2}{r_{21}^2} \hat{\mathbf{r}}_{21} \quad (1.10)$$

Some remarks on Eq. (1.10) are relevant:

- Eq. (1.10) is valid for any sign of q_1 , q_2 whether positive or negative. This is easily checked. If q_1 , q_2 are of the same sign (both positive or both negative), \mathbf{F}_{21} is along $\hat{\mathbf{r}}_{21}$, which denotes repulsion, as it should be for like charges. If q_1 , q_2 are of opposite signs, \mathbf{F}_{21} is along $-\hat{\mathbf{r}}_{21} = \hat{\mathbf{r}}_{12}$, which denotes attraction, as expected for unlike charges. Thus, we do not have to write separate equations for the cases of like and unlike charges. Eq. (1.10) takes care of both cases correctly (Fig. 1.5).
- Clearly from Eq. (1.10), \mathbf{F}_{12} is obtained by simply interchanging 1 and 2, i.e.,

$$\mathbf{F}_{12} = \frac{1}{4\pi\epsilon_0} \frac{q_1 q_2}{r_{12}^2} \hat{\mathbf{r}}_{12} = -\mathbf{F}_{21}$$

Thus, Coulomb's law agrees with Newton's third law.

- Coulomb's law [Eq. (1.10)] gives the force between two charges q_1 , q_2 in vacuum.

* Strictly, in SI units, 1 coulomb equals 1 ampere-second, where 1 ampere is defined in terms of the magnetic force between two current-carrying wires (Chapter 5).

If the charges are placed in matter or the intervening space has matter, the situation gets complicated due to the presence of charged constituents of matter. We shall consider electrostatics in matter in the next Chapter.

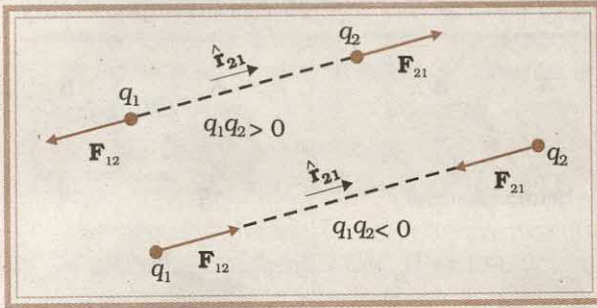


Fig. 1.5 Directions of attractive and repulsive Coulomb forces.

Example 1.1 A charged metallic sphere A is suspended by a nylon thread. Another charged metallic sphere B carried by an insulating handle is brought close to A to a distance of 9.0 cm between their centres. The resulting repulsion of A is noted (for example, by shining a beam of light and measuring the deflection of its shadow on a calibrated screen). Spheres A and B are touched by uncharged spheres C and D, respectively. C and D are then removed and B is brought closer to A to a distance of 4.5 cm between their centres. What is the expected repulsion of A on the basis of Coulomb's law? Spheres A and C and spheres B and D have identical sizes. Ignore the sizes of A and B in comparison to the separation between their centres.

Answer Let the original charge on sphere A be q and that on B be q' . At a distance r between their centres, the magnitude of the electrostatic force on each is given by

$$F = k \frac{qq'}{r^2}$$

neglecting the sizes of spheres A and B in comparison to r . When an identical but uncharged sphere C touches A, the charges redistribute on A and C and, by symmetry, each sphere carries a charge $(q/2)$. Similarly, after B touches D, the redistributed charge on each is $(q'/2)$. If now the separation between A and B is halved, the magnitude of the electrostatic

force on each is

$$F' = k \frac{(q/2)(q'/2)}{(r/2)^2} = k \frac{qq'}{r^2} = F$$

Thus the electrostatic force of repulsion on A remains unaltered.

Example 1.2 Coulomb's law for electrical force between two charges and Newton's law for gravitational force between two masses, both have inverse-square dependence on the distance between the charges/masses. Compare the strength of these forces by determining the ratio of their magnitude for an electron-proton system.

Answer The electric force between an electron and a proton at a distance r apart is :

$$F_e = \frac{-k e^2}{r^2}$$

where the negative sign indicates that the force is attractive. The corresponding gravitational force (always attractive) is :

$$F_g = -G \frac{m_p m_e}{r^2}$$

where m_p and m_e are the masses of the proton and electron.

$$\left| \frac{F_e}{F_g} \right| = \frac{k e^2}{G m_p m_e} = 2.4 \times 10^{39}$$

The (dimensionless) ratio of the two forces shows that electrical forces are enormously stronger than the gravitational forces. This is evident from common experience. When you hold a book in your hand, the electric forces between the palm of your hand and the book (why are there electric forces, even though both the book and the hand are not charged?) are strong enough to counter the gravitational force on the book due to the entire earth!

1.6 BASIC PROPERTIES OF ELECTRIC CHARGE

It will be useful at this point to note some important basic properties of electric charge:

Additivity

Consider a system of two point charges q_1 and q_2 . The total charge of the system is obtained simply

by adding q_1 and q_2 . Thus, charges add up like real numbers (scalars). The total charge of a system containing charges q_1, q_2, \dots, q_n is $q_1 + q_2 + \dots + q_n$. This is like the property of mass. Both mass and charge are scalars. They have no direction in space associated with them. Note, however, one difference. Mass is always a non-negative number. Charge can be positive or negative. When you add charges, you must take care of their signs also. The total charge of a system containing three charges $2 \mu\text{C}$, $-5 \mu\text{C}$ and $-6 \mu\text{C}$ is $-9 \mu\text{C}$. We have implicitly made use of the additive property of charge in the earlier discussion also.

Conservation of electric charge

We have mentioned this property before. Charge can be transferred from one body to another, but can neither be created nor destroyed. If in a given volume of space, you find that charge has increased with time, it is certain that some charge from outside has entered into the given volume of space. On the other hand, if it has decreased, some charge has exited out of the given volume of space. If a system is isolated, i.e., if no charge can get into or out of the system, the total electric charge of the system does not change with time. Within the isolated system, interactions between different bodies of the system can cause transfer of charge from one body to another, but the total charge of the isolated system is conserved.

The law of conservation of electric charge is an exact law of nature. We saw it in the context of frictional electricity in section 1.2. But it is true in all domains of nature. Even in high energy physics domains, where mass can be converted into energy and vice-versa (as predicted by Einstein's theory of relativity), the law of charge conservation continues to be strictly valid.

Quantisation of electric charge

The statement that a certain quantity is quantised occurs frequently in physics. It means that the quantity can take any one of only a discrete set of values. This is so about electric charge. It is found experimentally that electric charge of any system in nature is always an integer multiple of a certain lowest amount

of charge. Thus the charge q of a body is always given by

$$q = ne$$

where e (> 0) is the lowest possible magnitude of charge and n belongs to the set of integers: $n = 0, \pm 1, \pm 2, \pm 3, \dots$. A charge q equal to, say, $1.735 e$ or $-567.9 e$ or $\sqrt{2} e$ is an impossibility.

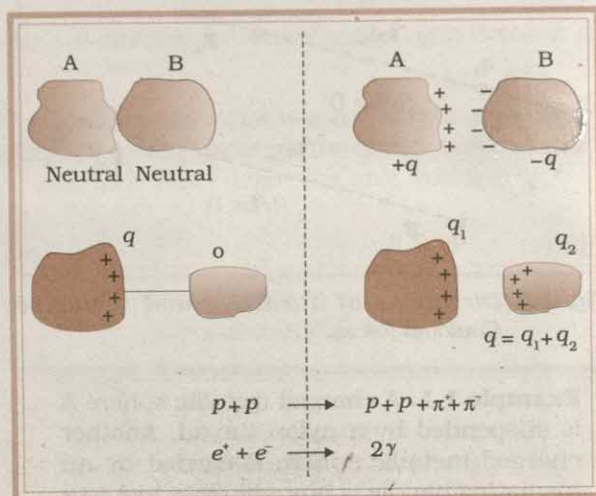


Fig. 1.6 Electric charge is additive. The electric charge of an isolated system is conserved. The left and right sides denote, respectively, the initial and final situations for some interaction. The last two processes are in high energy domain where mass is created (or destroyed), but the total charge remains unchanged.

We now know that the lowest possible charge e is in fact the magnitude of charge of an electron*. It is also equal to the charge of a proton.

The fact that electric charge is quantised was first suggested by the experimental laws of electrolysis discovered by Faraday. It was further established by the famous experiment due to Millikan in 1912 on the measurement of charge of an electron. There is no explanation in classical physics, or even in modern physics, why electric charge should be quantised. Quantisation of electric charge is an experimentally verified law in all domains of nature.

At the macroscopic level, one deals with charges that are enormous compared to the magnitude of charge e . Since $e = 1.6 \times 10^{-19} \text{ C}$, a charge of magnitude, say, $1 \mu\text{C}$ contains something like 10^{13} times the electronic charge.

*Recent discoveries in high energy physics have indicated that elementary constituents (called quarks) of proton, neutron, etc. have charges $(2/3)e$, $-(1/3)e$. We omit discussion of these new findings here (Chapter 14).

At this scale, the fact that charge can increase or decrease only in units of e is not very different from saying that charge can take continuous values. Thus, at the macroscopic level, the quantisation of charge has no practical consequence and can be ignored. At the microscopic level, where the charges involved are of the order of a few tens or hundreds of e , quantisation of charge cannot be ignored.

1.7 MULTIPLE CHARGES: THE SUPERPOSITION PRINCIPLE

The mutual electric force between two charges is given by Coulomb's law. How to calculate the force on a charge where there are not one but several charges around? Consider a system of n charges $q_1, q_2, q_3, \dots, q_n$. What is the force on q_1 due to q_2, q_3, \dots, q_n ?

Coulomb's law is not enough to answer this question. We need an additional principle, called the *Principle of Superposition*. What this says is that in a system of charges q_1, q_2, \dots, q_n , the force on q_1 due to q_2 is the same as given by Coulomb's law, i.e., it is unaffected by the presence of the other charges q_3, q_4, \dots, q_n . Thus, if the force on q_1 due to q_2 is denoted by \mathbf{F}_{12} , \mathbf{F}_{12} is given by Eq. (1.10), even though other charges are present.

$$\mathbf{F}_{12} = k \frac{q_1 q_2}{r_{12}^2} \hat{\mathbf{r}}_{12}$$

In the same way, the force on q_1 due to q_3 , denoted by \mathbf{F}_{13} , is given by

$$\mathbf{F}_{13} = k \frac{q_1 q_3}{r_{13}^2} \hat{\mathbf{r}}_{13}$$

which again is the Coulomb force on q_3 due to q_1 , even though other charges q_2, q_4, \dots, q_n are present.

The total force \mathbf{F}_1 on the charge q_1 is then given by the vector sum of the forces $\mathbf{F}_{12}, \mathbf{F}_{13}, \dots, \mathbf{F}_{1n}$.

$$\begin{aligned} \mathbf{F}_1 &= \mathbf{F}_{12} + \mathbf{F}_{13} + \dots + \mathbf{F}_{1n} \\ &= k \left[\frac{q_1 q_2}{r_{12}^2} \hat{\mathbf{r}}_{12} + \frac{q_1 q_3}{r_{13}^2} \hat{\mathbf{r}}_{13} + \dots \right] \quad (1.11) \end{aligned}$$

The vector sum is obtained as usual by the parallelogram law of vectors. All of electrostatics is basically a consequence of Coulomb's Law and the superposition principle.

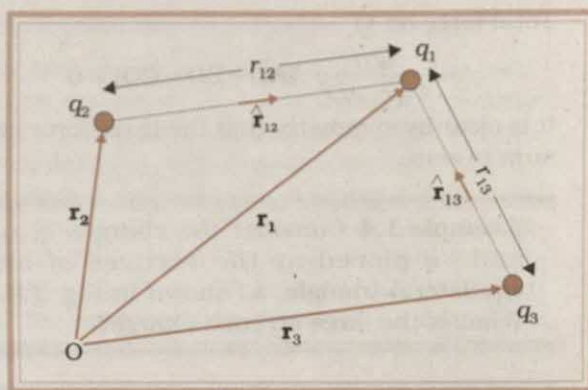


Fig. 1.7 Notations used in Eq. (1.11).

Example 1.3 Consider three charges q_1, q_2, q_3 each equal to q at the vertices of an equilateral triangle of side l . What is the force on a charge Q placed at the centroid of the triangle?

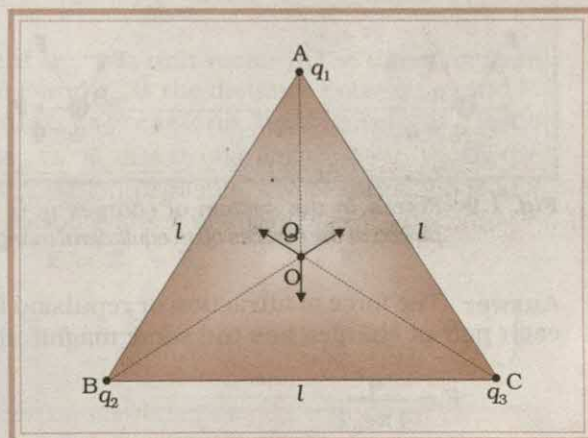


Fig. 1.8 Forces on the charge Q at the centroid of an equilateral triangle due to three equal charges q placed at its vertices. The directions of forces shown refer to the case when q and Q are of the same sign.

Answer

$$\text{Force on } Q \text{ due to } q_1 = \frac{1}{4\pi\epsilon_0} \frac{Qq_1}{AO^2} \hat{\mathbf{AO}}$$

$$\text{Force on } Q \text{ due to } q_2 = \frac{1}{4\pi\epsilon_0} \frac{Qq_2}{BO^2} \hat{\mathbf{BO}}$$

$$\text{Force on } Q \text{ due to } q_3 = \frac{1}{4\pi\epsilon_0} \frac{Qq_3}{CO^2} \hat{\mathbf{CO}}$$

Total force on Q

$$= \frac{Qq}{4\pi\epsilon_0 AO^2} (\hat{AO} + \hat{BO} + \hat{CO}) = 0$$

It is clear by symmetry that the three forces will sum to zero.

Example 1.4 Consider the charges q , q , and $-q$ placed at the vertices of an equilateral triangle, as shown in Fig. 1.9. What is the force on each charge?

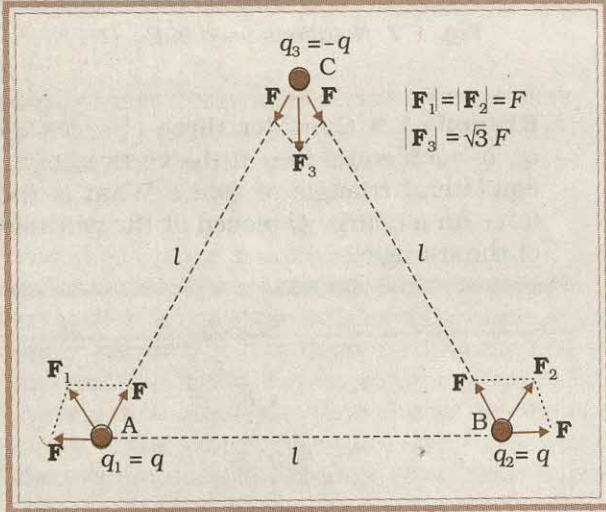


Fig. 1.9 Forces in the system of charges q , q , $-q$ placed at the vertices of an equilateral triangle.

Answer The force of attraction or repulsion for each pair of charges has the same magnitude

$$F = \frac{q^2}{4\pi\epsilon_0 l^2}$$

By the parallelogram law

$$\mathbf{F}_1 = F \hat{BC}$$

where \hat{BC} is a unit vector along BC;

$$\mathbf{F}_2 = F \hat{AC}$$

where \hat{AC} is a unit vector along AC;

$$\mathbf{F}_3 = \sqrt{3}F\mathbf{n}$$

where \mathbf{n} is the unit vector along the direction bisecting the angle (BCA).

It is interesting to see that the sum of the forces on the three charges is zero i.e.,

$$\mathbf{F}_1 + \mathbf{F}_2 + \mathbf{F}_3 = 0$$

The result is not at all surprising. It follows straight from the fact that Coulomb's law is consistent with Newton's third law. The proof is left to you as an exercise.

1.8 ELECTRIC FIELD

Let us return to Coulomb's law and consider the electrical force between two charges Q and q in vacuum. For simplicity, and without any loss of generality, let the charge Q be at the origin of the co-ordinate system. The force on q by Q is then given by

$$\mathbf{F} = k \frac{Qq}{R^2} \mathbf{R} \quad (1.12)$$

where \mathbf{R} is the unit vector in the direction from Q to q , and R is the distance between them. It is useful to look at the force \mathbf{F} arising in two steps:

- (i) The charge Q produces an electrical environment in the surrounding space, called an electric field. The electric field \mathbf{E} at any point in space due to a charge Q placed at the origin is given by

$$\mathbf{E} = k \frac{Q}{r^2} \mathbf{r} \quad (1.13)$$

where \mathbf{r} is the direction of the position vector from the origin to the point and r is the distance of the point from the origin.

- (ii) When a charge q is placed at a certain location $\mathbf{r} = \mathbf{R}$, without disturbing the location of Q , it experiences a force \mathbf{F} which equals the charge q multiplied by the electric field at the location of q . That is,

$$\mathbf{F} = q \mathbf{E}(\mathbf{R}) \quad (1.14)$$

Combining Eqs. (1.13) and (1.14) leads to Eq. (1.12). Clearly, the SI unit of electric field is N C^{-1} .

Some important remarks may be made here:

1. From Eq. (1.14), electric field due to a charge Q at a point in space may be defined as the force that a unit positive charge would experience if placed at that point. Note that the charge Q which is the source of the electric field must remain at its original location. (In general, the configuration of charges producing the electric field must remain undisturbed.) However, if a test charge q is brought at any point around Q , Q itself is bound to experience an electrical

force due to q and will tend to move. A way out of this difficulty is to make q negligibly small. The force \mathbf{F} is then negligibly small

but the ratio $\frac{\mathbf{F}}{q}$ is finite and defines the electric field:

$$\mathbf{E} = \lim_{q \rightarrow 0} \frac{\mathbf{F}}{q} \quad (1.15)^*$$

A practical way to get around the problem (of keeping Q undisturbed in the presence of q) is to hold Q to its location by unspecified forces! This may look strange but actually this is what happens in practice. When we are considering the electric force on a test charge q due to a charged planar sheet (Section 1.15), the charges on the sheet are held to their locations by the forces due to the unspecified charged constituents inside the sheet.

- Note that the electric field \mathbf{E} due to Q though defined operationally in terms of some test charge q , is independent of q . (This is because \mathbf{F} is proportional to q , so the ratio \mathbf{F}/q does not depend on q .) Further, while the force \mathbf{F} refers to a charge q at a particular location \mathbf{R} , the electric field \mathbf{E} due to Q is defined all over space. That is why, we have distinguished between \mathbf{r} , the general position vector, from the particular location $\mathbf{r} = \mathbf{R}$.

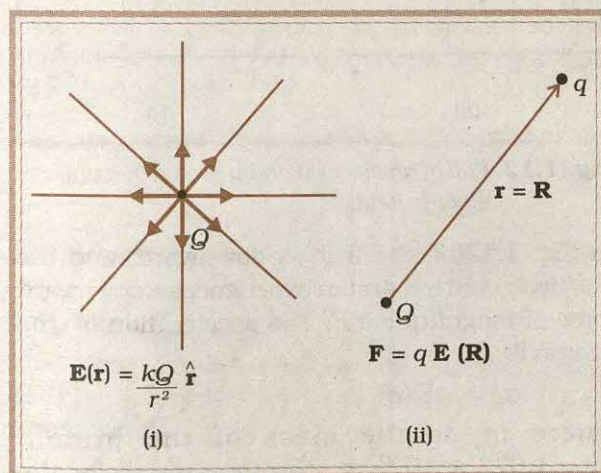


Fig. 1.10 Introducing the notion of electric field
(i) Charge Q produces an electric field $\mathbf{E}(\mathbf{r})$ all over space (ii) A charge q placed at \mathbf{R} experiences a force $q\mathbf{E}(\mathbf{R})$.

Electric field due to a system of charges

Consider a system of charges q_1, q_2, \dots, q_n with position vectors $\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_n$ relative to some origin O . Like the field due to a single charge, electric field at a point in space due to the system of charges is defined to be the force experienced by a unit test charge placed at that point, without disturbing the original positions of charges q_1, q_2, \dots, q_n . We can use Coulomb's law and the superposition principle to determine this field:

Electric field \mathbf{E}_1 at \mathbf{r} due to q_1 at \mathbf{r}_1 is given by

$$\mathbf{E}_1 = k \frac{q_1 \times 1}{r_{1P}^2} \mathbf{r}_{1P}$$

where \mathbf{r}_{1P} is a unit vector in the direction from q_1 to P , and r_{1P} is the distance between q_1 and P . In the same manner, electric field \mathbf{E}_2 at \mathbf{r} due to q_2 at \mathbf{r}_2 is

$$\mathbf{E}_2 = k \frac{q_2 \times 1}{r_{2P}^2} \mathbf{r}_{2P}$$

where \mathbf{r}_{2P} is a unit vector in the direction from q_2 to P and r_{2P} is the distance between q_2 and P . Similar expressions hold good for fields $\mathbf{E}_3, \mathbf{E}_4, \dots, \mathbf{E}_n$ due to charges q_3, q_4, \dots, q_n . By the superposition principle, the electric field \mathbf{E} at \mathbf{r} due to the system of charges is

$$\mathbf{E} = \mathbf{E}_1 + \mathbf{E}_2 + \dots + \mathbf{E}_n$$

$$= k \left[\frac{q_1}{r_{1P}^2} \mathbf{r}_{1P} + \frac{q_2}{r_{2P}^2} \mathbf{r}_{2P} + \dots + \frac{q_n}{r_{nP}^2} \mathbf{r}_{nP} \right] \quad (1.16)$$

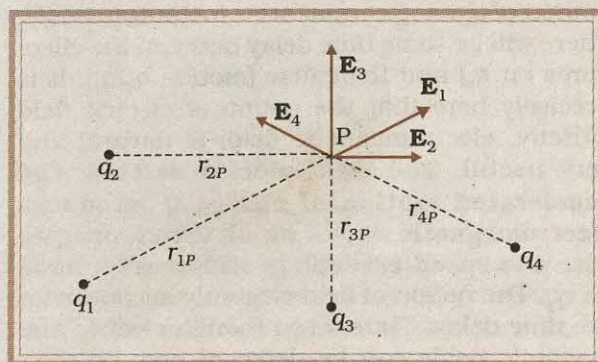


Fig. 1.11 Electric field at a point due to a system of charges is the vector sum of the electric fields at the point due to individual charges.

* Because of charge quantisation, the test charge q cannot go below e . However, on the macroscopic scale, this is as good as taking the limit $q \rightarrow 0$.

Physical significance of electric field

You may wonder why the notion of electric field has been introduced here at all. After all, for any system of charges, the measurable quantity is the force on a charge which can be directly determined using Coulomb's law and the superposition principle [Eq. (1.11)]. Why then introduce this intermediate quantity called the electric field?

For electrostatics, the concept of electric field is convenient, but not really necessary. Electric field is an elegant way of characterising the electrical environment of a system of charges. Electric field at a point in the space around a system of charges tells you the force a unit positive test charge would experience if placed at that point (without disturbing the system). Electric field is a characteristic of the system of charges and is independent of the test charge that you place at a point to determine the field. The term 'field' in physics generally refers to a quantity that is defined at every point in space and may vary from point to point. Electric field is a vector field, since force is a vector quantity.

The true physical significance of the concept of electric field, however, emerges only when we go beyond electrostatics and deal with time-dependent electromagnetic phenomena. Suppose we consider the force between two distant charges q_1 , q_2 in accelerated motion. Now the greatest speed with which a signal or information can go from one point to another is c , the speed of light. Thus, the effect of any motion of q_1 on q_2 cannot arise instantaneously. There will be some time delay between the effect (force on q_2) and the cause (motion of q_1). It is precisely here that the notion of electric field (strictly, electromagnetic field) is natural and very useful. **The field picture is this: the accelerated motion of charge q_1 produces electromagnetic waves, which then propagate with the speed c , reach q_2 and cause a force on q_2 .** The notion of field elegantly accounts for the time delay. Thus, even though electric and magnetic fields can be detected only by their effects (forces) on charges, they are regarded as physical entities, not merely mathematical constructs. The concept of field was first introduced by Faraday and is now among the central concepts in physics.

Example 1.5 An electron falls through a distance of 1.5 cm in a uniform electric field of magnitude $2.0 \times 10^4 \text{ N C}^{-1}$ [Fig. 1.12(a)]. The direction of the field is reversed keeping its magnitude unchanged and a proton falls through the same distance [Fig. 1.12(b)]. Compute the time of fall in each case. Contrast the situation with that of 'free fall under gravity'.

Answer In Fig. 1.12(a) the field is upward, so the negatively charged electron experiences a downward force of magnitude eE where E is the magnitude of the electric field. The acceleration of the electron is

$$a_e = eE/m_e$$

where m_e is the mass of the electron.

Starting from rest, the time required by the electron to fall through a distance h is given by

$$t_e = \sqrt{\frac{2h}{a_e}} = \sqrt{\frac{2hm_e}{eE}}$$

For $e = 1.602 \times 10^{-19} \text{ C}$, $m_e = 9.110 \times 10^{-31} \text{ kg}$,
 $E = 2.0 \times 10^4 \text{ N C}^{-1}$, $h = 1.5 \times 10^{-2} \text{ m}$,

$$t_e = 2.9 \times 10^{-9} \text{ s}.$$

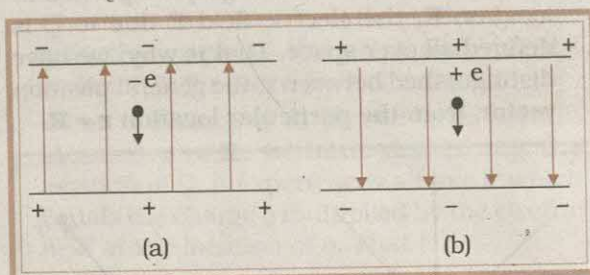


Fig. 1.12 Fall of an electron and a proton in a uniform electric field.

In Fig. 1.12(b), the field is downward, and the positively charged proton experiences a downward force of magnitude eE . The acceleration of the proton is

$$a_p = eE/m_p$$

where m_p is the mass of the proton; $m_p = 1.673 \times 10^{-27} \text{ kg}$. The time of fall for the proton is

$$t_p = \sqrt{\frac{2h}{a_p}} = \sqrt{\frac{2hm_p}{eE}} = 1.3 \times 10^{-7} \text{ s}.$$

Thus, the heavier particle (proton) takes a greater time to fall through the same distance.

This is in basic contrast to the situation of 'free fall under gravity' where the time of fall is independent of the mass of the body. Note that in this example we have ignored the acceleration due to gravity in calculating the time of fall. To see if this is justified, let us calculate.

$$a_p = \frac{eE}{m_p}$$

$$= \frac{(1.602 \times 10^{-19} \text{ C}) \times (2.0 \times 10^4 \text{ N C}^{-1})}{1.673 \times 10^{-27} \text{ kg}}$$

$$= 1.9 \times 10^{12} \text{ m s}^{-2}$$

which is enormous compared to the value of g (9.8 m s^{-2}). The acceleration of the electron is even greater. Thus, the effect of acceleration due to gravity can be ignored in this example. ◀

Example 1.6 Two point charges q_1 and q_2 of $+10^{-8} \text{ C}$ and -10^{-8} C , respectively, are placed 0.1 m apart. Calculate the electric fields at points A, B and C shown in Fig. 1.13.

Answer The electric field vector \mathbf{E}_1 at A due to the positive charge q_1 points towards the right and has a magnitude

$$E_1 = \frac{(9 \times 10^9 \text{ N m}^2 \text{ C}^{-2}) \times (10^{-8} \text{ C})}{(0.05 \text{ m})^2}$$

$$= 3.6 \times 10^4 \text{ N C}^{-1}.$$

The electric field vector \mathbf{E}_2 at A due to the negative charge q_2 points towards the right and has a magnitude

$$E_2 = \frac{(9 \times 10^9 \text{ N m}^2 \text{ C}^{-2}) \times (10^{-8} \text{ C})}{(0.05 \text{ m})^2}$$

$$= 3.6 \times 10^4 \text{ N C}^{-1}.$$

The magnitude of the total electric field at A is

$$E_A = E_1 + E_2 = 7.2 \times 10^4 \text{ N C}^{-1}$$

\mathbf{E}_A is directed toward the right.

The electric field vector \mathbf{E}_1 at B due to the positive charge q_1 points towards the left and has a magnitude

$$E_1 = \frac{(9 \times 10^9 \text{ N m}^2 \text{ C}^{-2}) \times (10^{-8} \text{ C})}{(0.05 \text{ m})^2}$$

$$= 3.6 \times 10^4 \text{ N C}^{-1}.$$

The electric field vector \mathbf{E}_2 at B due to the negative charge q_2 points towards the right and has a magnitude

$$E_2 = \frac{(9 \times 10^9 \text{ N m}^2 \text{ C}^{-2}) \times (10^{-8} \text{ C})}{(0.15 \text{ m})^2}$$

$$= 4 \times 10^3 \text{ N C}^{-1}.$$

The magnitude of the total electric field at B is

$$E_B = 3.2 \times 10^4 \text{ N C}^{-1}$$

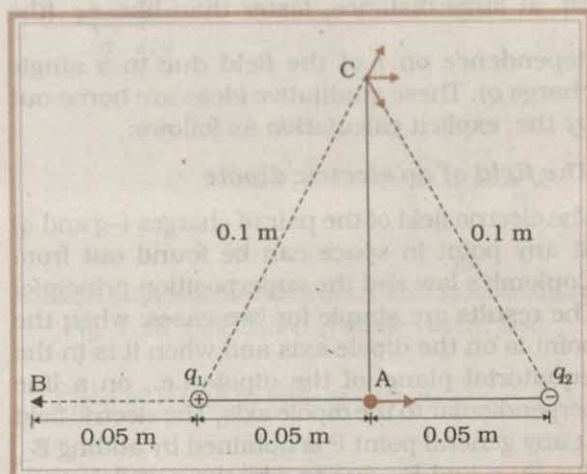


Fig. 1.13 Calculation of total electric field due to a system of two charges.

\mathbf{E}_B is directed towards the left. The magnitude of each electric field vector at point C, due to charge q_1 and q_2 is:

$$E_1 = E_2 = \frac{(9 \times 10^9 \text{ N m}^2 \text{ C}^{-2}) \times (10^{-8} \text{ C})}{(0.10 \text{ m})^2}$$

$$= 9 \times 10^3 \text{ N C}^{-1}.$$

The directions in which these two vectors point are indicated in the figure. The resultant of these two vectors is

$$E_C = E_1 \cos \frac{\pi}{3} + E_2 \cos \frac{\pi}{3}$$

$$= 9 \times 10^3 \text{ N C}^{-1}$$

\mathbf{E}_C points towards the right. ▶

1.9 ELECTRIC DIPOLE

An electric dipole is a pair of equal and opposite point charges q and $-q$ separated by a distance $2a$. The line connecting the two charges defines a direction in space. By convention, the direction from $-q$ to q is said to be the axis of the dipole.

The total charge of the electric dipole is obviously zero. This does not mean that the field of the electric dipole is zero. Since the charge q and $-q$ are separated by some distance, the

electric fields due to them when added do not exactly cancel out. However, at distances much larger than the separation ($r \gg 2a$), the fields due to q and $-q$ do nearly cancel out. We thus expect that the electric field due to a dipole falls

off, at large distance, faster than like $\frac{1}{r^2}$ (the dependence on r of the field due to a single charge q). These qualitative ideas are borne out by the explicit calculation as follows:

The field of an electric dipole

The electric field of the pair of charges ($-q$ and q) at any point in space can be found out from Coulomb's law and the superposition principle. The results are simple for two cases: when the point is on the dipole axis and when it is in the 'equatorial plane' of the dipole i.e., on a line perpendicular to the dipole axis. The electric field at any general point P is obtained by adding \mathbf{E}_{-q} due to $-q$ and \mathbf{E}_{+q} due to q by the parallelogram law of vectors.

For points on the axis

Let the point be at distance r from the centre of the dipole on the side of the charge q . Then

$$\mathbf{E}_{-q} = -\frac{q}{4\pi\epsilon_0(r+a)^2} \hat{\mathbf{p}} \quad (1.17a)$$

where $\hat{\mathbf{p}}$ is the unit vector along the dipole axis (from $-q$ to q). Also

$$\mathbf{E}_{+q} = \frac{q}{4\pi\epsilon_0(r-a)^2} \hat{\mathbf{p}} \quad (1.17b)$$

The total field at P is

$$\begin{aligned} \mathbf{E} &= \mathbf{E}_{+q} + \mathbf{E}_{-q} = \frac{q}{4\pi\epsilon_0} \left[\frac{1}{(r-a)^2} - \frac{1}{(r+a)^2} \right] \hat{\mathbf{p}} \\ &= \frac{q}{4\pi\epsilon_0} \frac{4ar}{(r^2 - a^2)^2} \hat{\mathbf{p}} \end{aligned} \quad (1.18)$$

For $r \gg a$

$$\mathbf{E} = \frac{4qa}{4\pi\epsilon_0 r^3} \hat{\mathbf{p}} \quad (r \gg a) \quad (1.19)$$

For points on the equatorial plane

$$\mathbf{E}_{+q} = \frac{q}{4\pi\epsilon_0} \frac{1}{r^2 + a^2} \quad (1.20)$$

$$\mathbf{E}_{-q} = \frac{q}{4\pi\epsilon_0} \frac{1}{r^2 + a^2} \quad (1.21)$$

The directions of \mathbf{E}_{+q} and \mathbf{E}_{-q} are as shown in Fig. 1.14 (b). Clearly, the components normal to the dipole axis cancel away. The components along the dipole axis add up. The total electric field is opposite to $\hat{\mathbf{p}}$:

$$\begin{aligned} \mathbf{E} &= -(\mathbf{E}_{+q} + \mathbf{E}_{-q}) \cos\theta \hat{\mathbf{p}} \\ &= -\frac{2q}{4\pi\epsilon_0} \frac{1}{(r^2 + a^2)} \frac{a}{(r^2 + a^2)^{1/2}} \hat{\mathbf{p}} \\ &= -\frac{2qa}{4\pi\epsilon_0 (r^2 + a^2)^{3/2}} \hat{\mathbf{p}} \end{aligned} \quad (1.22)$$

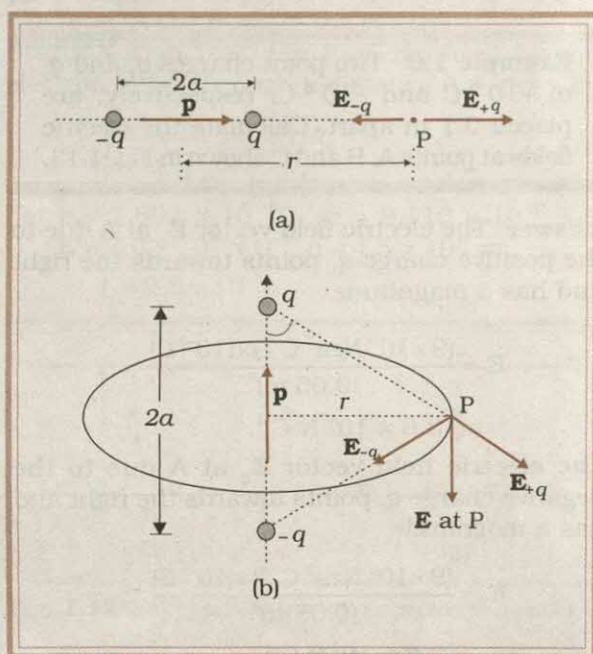


Fig. 1.14 Electric field of a dipole at (a) a point on the axis (b) a point on the equatorial plane of the dipole. \mathbf{p} is the dipole moment vector of magnitude $p = q \times 2a$ and direction from $-q$ to q .

At large distances ($r \gg a$), this reduces to

$$\mathbf{E} = -\frac{2qa}{4\pi\epsilon_0 r^3} \hat{\mathbf{p}} \quad (r \gg a) \quad (1.23)$$

From Eqs. (1.19) and (1.23), it is clear that the dipole field at large distances does not involve q and a separately; it depends on the combination qa . This suggests the definition of dipole moment.

The dipole moment vector \mathbf{p} of an electric dipole is defined by

$$\mathbf{p} = q \times 2a \hat{\mathbf{p}} \quad (1.24)$$

i.e., it is a vector whose magnitude is charge q times the separation $2a$ (between the pair of charges $q, -q$) and the direction is along the line from $-q$ to q . In terms of \mathbf{p} , the electric field of a dipole at large distance takes the simple forms :

At a point on the dipole axis

$$\mathbf{E} = \frac{2\mathbf{p}}{4\pi\epsilon_0 r^3} \quad (r \gg a) \quad (1.25)$$

At a point on the equatorial plane

$$\mathbf{E} = -\frac{\mathbf{p}}{4\pi\epsilon_0 r^3} \quad (r \gg a) \quad (1.26)$$

Notice the important point that the dipole field

at large distances falls off not as $\frac{1}{r^2}$ but as $\frac{1}{r^3}$.

Further, the magnitude and direction of the dipole field depends not only on the distance r but also on the angle between the position vector \mathbf{r} and the dipole moment \mathbf{p} .

We can think of the limit when the dipole size $2a$ approaches zero, the charge q approaches infinity in such a way that the product $p = q \times 2a$ is finite. Such a dipole is referred to as a **point dipole**. For a point dipole, Eqs. (1.25) and (1.26) are exact, true for any r .

Example 1.7 Two charges $\pm 10 \mu\text{C}$ are placed 5.0 mm apart. Determine the electric field at (a) a point P on the axis of the dipole 15 cm away from its centre O on the side of the positive charge, (b) a point Q, 15 cm away from O on a line passing through O and normal to the axis of the dipole.

Answer

(a) Field at P due to charge $+10 \mu\text{C}$

$$\begin{aligned} &= \frac{10^{-5} \text{ C}}{4\pi(8.854 \times 10^{-12} \text{ C}^2 \text{ N}^{-1} \text{ m}^{-2})} \\ &\times \frac{1}{(15 - 0.25)^2 \times 10^{-4} \text{ m}^2} \\ &= 4.13 \times 10^6 \text{ N C}^{-1} \text{ along BP.} \end{aligned}$$

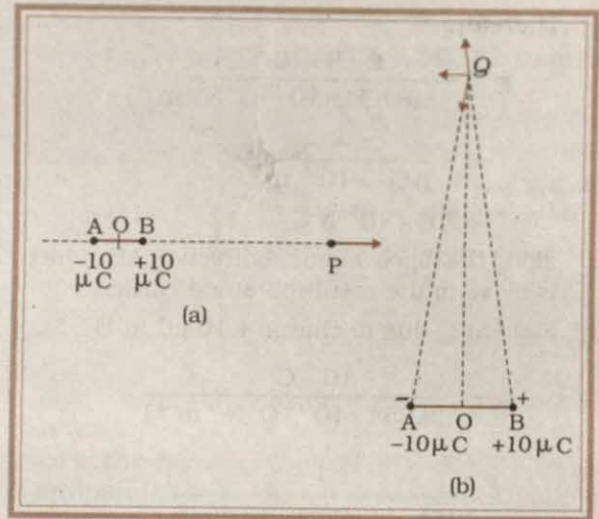


Fig. 1.15 Calculation of dipole electric field.

Field at P due to charge $-10 \mu\text{C}$

$$\begin{aligned} &= \frac{10^{-5} \text{ C}}{4\pi(8.854 \times 10^{-12} \text{ C}^2 \text{ N}^{-1} \text{ m}^{-2})} \\ &\times \frac{1}{(15 + 0.25)^2 \times 10^{-4} \text{ m}^2} \\ &= 3.86 \times 10^6 \text{ N C}^{-1} \text{ along PA.} \end{aligned}$$

The resultant electric field at P due to the two charges at A and B is

$$= 2.7 \times 10^5 \text{ N C}^{-1} \text{ along BP.}$$

In this example, the ratio OP/OB is quite large ($=60$). Thus, we can expect to get approximately the same result as above by directly using the formula for electric field at a far-away point on the axis of a dipole. For a dipole consisting of charges $\pm q$, $2a$ distance apart, the electric field at a distance r from the centre on the axis of the dipole has a magnitude

$$E = \frac{2p}{4\pi\epsilon_0 r^3} \quad (r/a \gg 1)$$

where $p = 2aq$ is the magnitude of the dipole moment.

The direction of electric field on the dipole axis is always along the direction of the dipole moment vector (i.e., from $-q$ to q). Here

$$p = 10^{-5} \text{ C} \times 5 \times 10^{-3} \text{ m} = 5 \times 10^{-8} \text{ C m}$$

Therefore,

$$E = \frac{2 \times 5 \times 10^{-8} \text{ C m}}{4\pi(8.854 \times 10^{-12} \text{ C}^2 \text{ N}^{-1} \text{ m}^{-2})} \times \frac{1}{(15)^3 \times 10^{-6} \text{ m}^3}$$

$$= 2.6 \times 10^5 \text{ N C}^{-1}$$

along the dipole moment direction AB, which is close to the result obtained earlier.

(b) Field at Q due to charge + 10 μC at B

$$= \frac{10^{-5} \text{ C}}{4\pi(8.854 \times 10^{-12} \text{ C}^2 \text{ N}^{-1} \text{ m}^{-2})} \times \frac{1}{[15^2 + (0.25)^2] \times 10^{-4} \text{ m}^2}$$

$$= 3.99 \times 10^6 \text{ N C}^{-1} \text{ along BQ.}$$

Field at Q due to charge -10 μC at A

$$= \frac{10^{-5} \text{ C}}{4\pi(8.854 \times 10^{-12} \text{ C}^2 \text{ N}^{-1} \text{ m}^{-2})} \times \frac{1}{[15^2 + (0.25)^2] \times 10^{-4} \text{ m}^2}$$

$$= 3.99 \times 10^6 \text{ N C}^{-1} \text{ along QA.}$$

Clearly, the components of these two forces with equal magnitudes cancel along the direction OQ but add up along the direction parallel to BA. Therefore, the resultant electric field at Q due to the two charges at A and B is

$$= 2 \times \frac{0.25}{\sqrt{15^2 + (0.25)^2}} \times 3.99 \times 10^6 \text{ N C}^{-1}$$

along BA


$$= 1.33 \times 10^5 \text{ N C}^{-1} \text{ along BA.}$$

As in (a), we can expect to get approximately the same result by directly using the formula for dipole field at a point on the normal to the axis of the dipole:

$$E = \frac{p}{4\pi\epsilon_0 r^3} \quad (r/a \gg 1)$$

$$= \frac{5 \times 10^{-8} \text{ C m}}{4\pi(8.854 \times 10^{-12} \text{ C}^2 \text{ N}^{-1} \text{ m}^{-2})} \times \frac{1}{(15)^3 \times 10^{-6} \text{ m}^3}$$

$$= 1.33 \times 10^5 \text{ N C}^{-1}$$

The direction of electric field in this case is opposite to the direction of the dipole moment vector. Again the result agrees with that obtained before. 

Physical significance of dipoles

The study of electric dipoles is especially important for electrical phenomena in matter. Ordinarily, matter consists of neutral atoms or molecules. In a molecule, there are the positively charged nuclei and the negatively charged electrons. If the centre of mass of the positive charge does not coincide with that of the negative charge, the molecule has intrinsic (or permanent) dipole moment. Such molecules are called polar molecules. When there is no external field, the dipole moments of different molecules in a piece of matter are randomly oriented, so there is no net total dipole moment. In the presence of an external field, the polar molecules tend to align with the field and a net dipole moment results. We say that the matter has got 'polarised'.

Polarisation under an external field is possible even if the molecules of the matter are non-polar. Non-polar molecules, in the absence of field, have zero dipole moment. But in the presence of an external electric field, the negative charges and the positive charges in the molecule get displaced in opposite directions. That is, the atom or molecule, though neutral, has its centres of positive and negative charges slightly separated. Thus, the external field induces a dipole moment in the molecule in the direction of the field. The induced dipole moments of different molecules of the sample of matter add up resulting in a net total dipole moment.

The effect of polarisation is considered in some detail in Chapter 2.

1.10 DIPOLE IN A UNIFORM EXTERNAL FIELD

Consider a permanent dipole of dipole moment \mathbf{p} in a uniform external field \mathbf{E} , as shown in Fig. 1.16. (By permanent dipole, we mean that \mathbf{p} exists independent of \mathbf{E} ; it has not been induced by \mathbf{E} .)

There is a force $q\mathbf{E}$ on q and a force $-q\mathbf{E}$ on $-q$. The net force on the dipole is zero, since \mathbf{E} is uniform. However, the charges are separated, so the forces act at different points, resulting in a torque on the dipole. When the net force is

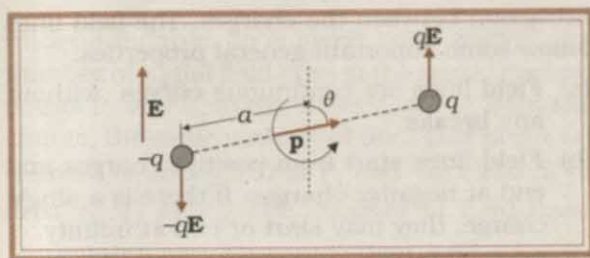


Fig. 1.16 Dipole in an external field.

zero, the torque (couple) is independent of the origin. Its magnitude equals the magnitude of each force multiplied by the arm of the couple (perpendicular distance between the two antiparallel forces):

$$\begin{aligned}\text{Magnitude of torque} &= qE \times 2a \sin\theta \\ &= 2qaE \sin\theta\end{aligned}$$

Its direction is normal to the plane of the paper, coming out of it.

The magnitude of $\mathbf{p} \times \mathbf{E}$ is also $pE \sin\theta$ and its direction is normal to the paper, coming out of it. Thus

$$\boldsymbol{\tau} = \mathbf{p} \times \mathbf{E} \quad (1.27)$$

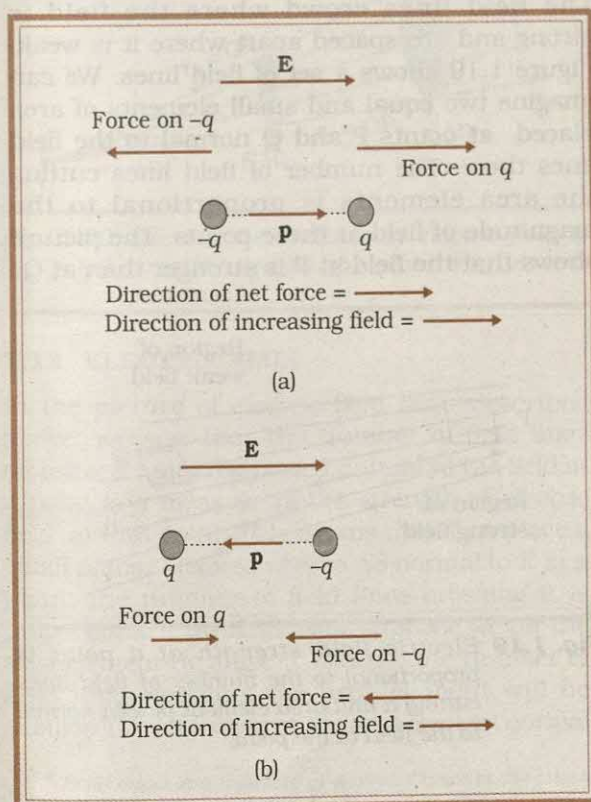


Fig. 1.17 Forces on a dipole in a non-uniform electric field.

This torque will tend to align the dipole with the field \mathbf{E} . When \mathbf{p} is aligned with \mathbf{E} , the torque is zero.

What happens if the field is not uniform? In that case, the net force will evidently be non-zero. In addition there will, in general, be a torque on the system as before. The general case is involved, so let us consider the simpler situations when \mathbf{p} is parallel to \mathbf{E} or antiparallel to \mathbf{E} . In either case, the net torque is zero, but there is a net force on the dipole if \mathbf{E} is not uniform.

Fig. 1.17 is self-explanatory. It is easily seen that when \mathbf{p} is parallel to \mathbf{E} , the dipole has a net force in the direction of increasing field. When \mathbf{p} is antiparallel to \mathbf{E} , the net force on the dipole is in the direction of decreasing field. In general, the force depends on the orientation of \mathbf{p} with respect to \mathbf{E} .

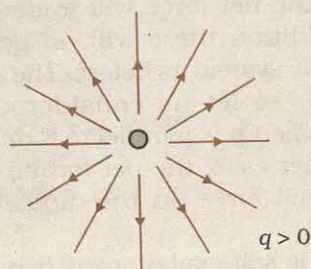
This brings us to a common observation in frictional electricity. A comb run through dry hair attracts pieces of paper. The comb, as we know, acquires charge through friction. But the paper is not charged. What then explains the attractive force? Taking the cue from the preceding discussion, the charged comb 'polarises' the piece of paper i.e., induces a net dipole moment in the direction of field. Further, the electric field due to the comb is not uniform. In this situation, it is easily seen that the paper should move in the direction of the comb!

1.11 ELECTRIC FIELD LINES

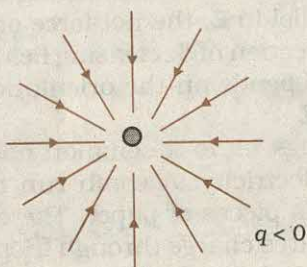
Electric field lines, invented by Faraday, are a way of pictorially mapping the electric field around a configuration of charges. An electric field line is, in general, a curve drawn in such a way that the tangent to it at each point is in the direction of the net field at that point. An arrow on the curve is obviously necessary to specify the direction of electric field from the two possible directions indicated by a tangent to the curve. A field line is a space curve, i.e., a curve in three dimensions.

Figure 1.18 shows the field lines around some simple charge configurations. The field lines are in 3-dimensional space, though the figure shows them only in a plane. The field lines of a single positive charge are radially outward while those of a single negative charge are radially inward. The field lines around a system of two positive charges (q, q) give a vivid pictorial description of

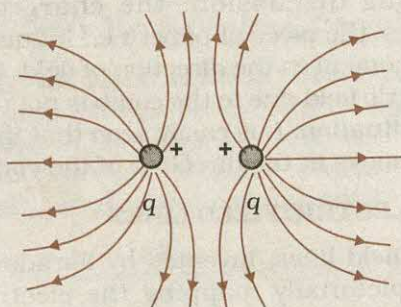
their mutual repulsion, while those around the configuration of two equal and opposite charges ($+q$, $-q$), a dipole, show clearly the mutual



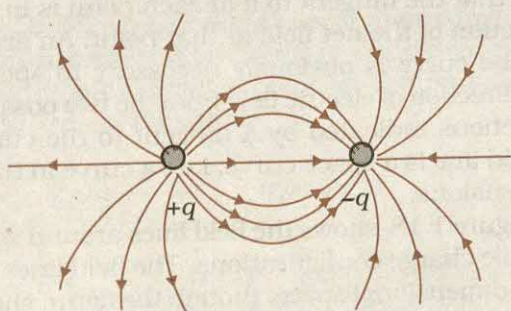
(a)



(b)



(c)



(d)

Fig. 1.18 Electrostatic field lines around some simple charge configurations.

attraction between the charges. The field lines follow some important general properties:

- Field lines are continuous curves without any breaks.
- Field lines start from positive charges and end at negative charges. If there is a single charge, they may start or end at infinity.
- Two field lines can never cross each other. (If they did, the field at the point of intersection will not have a unique direction, which is absurd.)
- Electrostatic field lines do not form any closed loops. This follows from the conservative nature of electric field (Chapter 2).

The field lines carry information about the direction of electric field at different points in space. What about the strength or magnitude of electric field? Having drawn a certain set of field lines, the relative density (i.e., closeness) of the field lines at different points indicates the relative strength of electric field at those points. The field lines crowd where the field is strong and are spaced apart where it is weak. Figure 1.19 shows a set of field lines. We can imagine two equal and small elements of area placed at points P and Q normal to the field lines there. The number of field lines cutting the area elements is proportional to the magnitude of field at these points. The picture shows that the field at P is stronger than at Q.

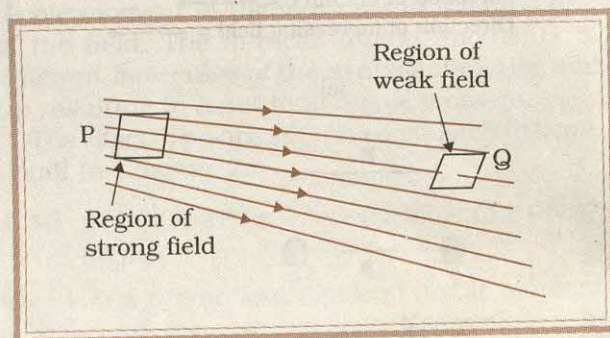


Fig. 1.19 Electric field strength at a point is proportional to the number of field lines cutting a unit area element placed normal to the field at the point.

Using this picture, the $\frac{1}{r^2}$ dependence of electric field due to a single charge q seems

understandable. In a given solid angle* the number of radial field lines is the same. For two points P_1 and P_2 at distances r_1 and r_2 from the charge, the same number of lines (say n) cut an element of area $r_1^2 \Delta\Omega$ at P_1 and an element of area $r_2^2 \Delta\Omega$ at P_2 . The number of field lines cutting unit area element is then $n/r_1^2 \Delta\Omega$ at P_1 and $n/r_2^2 \Delta\Omega$ at P_2 . Since n and $\Delta\Omega$ are common, the field clearly has a $\frac{1}{r^2}$ dependence.

The picture of field lines was invented by Faraday to develop an intuitive non-mathematical way of visualising electric fields around charged configurations. Faraday called them 'lines of force'. This term is somewhat misleading, especially in case of magnetic fields. The more appropriate term is 'field lines' (electric or magnetic) that we have adopted in this book.

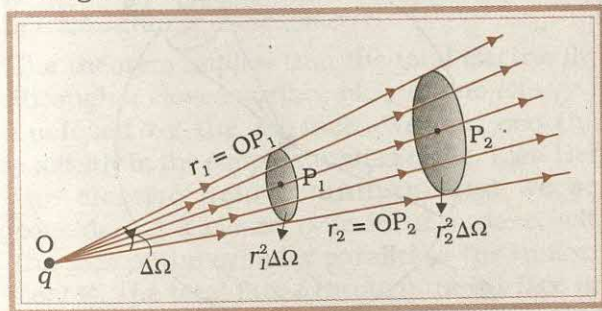


Fig. 1.20 The $1/r^2$ dependence of electric field is consistent with the picture of electric field lines.

1.12 ELECTRIC FLUX

In the picture of electric field lines described above, we saw that the number of field lines crossing a unit area placed normal to the field at a point is a measure of the strength of electric field at that point. This means that if we place a small planar element of area ΔS normal to \mathbf{E} at a point, the number of field lines crossing it is proportional** to $E \Delta S$. Now suppose we tilt the area element by angle θ . Clearly, the number of field lines crossing the area element will be smaller. The projection of the area element normal

to \mathbf{E} is $\Delta S \cos\theta$. Thus, the number of field lines crossing ΔS is proportional to $E \Delta S \cos\theta$. When $\theta = 90^\circ$, field lines will be parallel to ΔS and not cross it at all (Fig. 1.21).

The orientation of area element and not merely its magnitude is important in many contexts.

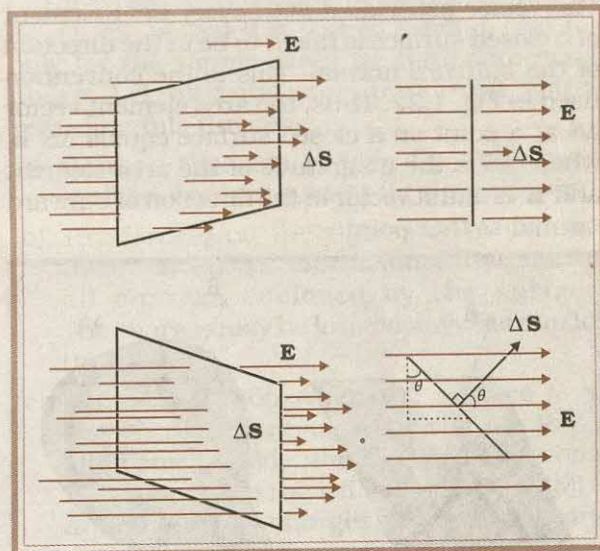


Fig. 1.21 The number of field lines crossing a small area element ΔS at a point depends on the orientation of the area relative to the electric field direction. The right side of the figure gives a cross-sectional view.

For example, in a stream, the amount of water flowing through a ring will naturally depend on how you hold the ring. If you hold it normal to the flow, more water will flow through it than if you hold it with some other orientation. This shows that an area element should be treated as a vector. It has a magnitude as also a direction. How to specify the direction of a planar area? Clearly, the normal to the plane specifies the orientation of the plane. Thus the direction of a planar area vector is along its normal.

How to associate a vector to the area of a curved surface? We imagine dividing the surface into a large number of very small area elements. Each small area element may be treated as planar and a vector associated with it as explained before.

Notice one ambiguity here. The direction of an area element is along its normal. But a

* Solid angle is a measure of a cone. Consider the intersection of the given cone with a sphere of radius R . The solid angle $\Delta\Omega$ of the cone is defined to be equal to $\Delta S/R^2$, where ΔS is the area on the sphere cut out by the cone.

** It will not be proper to say that the number of field lines is equal to $E \Delta S$. The number of field lines is after all, a matter of how many field lines we choose to draw. What is physically significant is the relative number of field lines crossing a given area at different points.

normal can point in two directions. Which direction do we choose as the direction of the vector associated with the area element? This problem is resolved by some convention appropriate to the given context. For the case of a closed surface, this convention is very simple. The vector associated with every area element of a closed surface is taken to be in the direction of the **outward** normal. This is the convention used in Fig. 1.22. Thus, the area element vector $\Delta \mathbf{S}$ at a point on a closed surface equals $\Delta S \mathbf{n}$ where ΔS is the magnitude of the area element and \mathbf{n} is a unit vector in the direction of outward normal at that point.

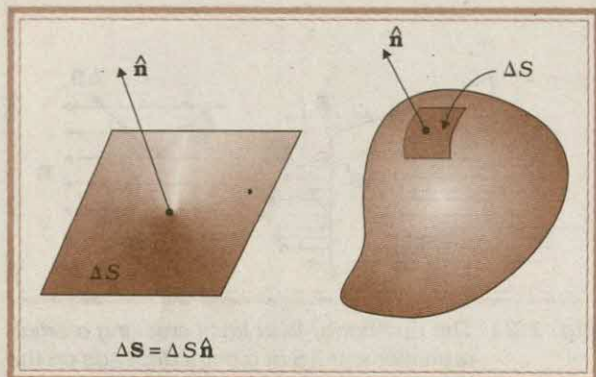


Fig. 1.22 Area is a vector quantity because not only its magnitude but also its orientation is significant. The orientation of a planar area is specified by its normal. Every small element of a curved surface can be treated as a planar area.

We now come to the definition of electric flux. Electric flux $\Delta \phi$ through an area element $\Delta \mathbf{S}$ is defined by

$$\Delta \phi \doteq \mathbf{E} \cdot \Delta \mathbf{S} = E \Delta S \cos \theta \quad (1.28)$$

which, as seen before, is proportional to the number of field lines cutting the area element. The angle θ here is the angle between \mathbf{E} and $\Delta \mathbf{S}$. For a closed surface, with the convention stated already, θ is the angle between \mathbf{E} and the outward normal \mathbf{n} to the area element. Notice we could look at the expression $E \Delta S \cos \theta$ in two ways: $E (\Delta S \cos \theta)$ i.e., E times the projection of area normal to \mathbf{E} , or $E_{\perp} \Delta S$ i.e., component of \mathbf{E} along the normal to the area element times the magnitude of the area element. The unit of electric flux is $\text{N C}^{-1} \text{m}^2$.

The basic definition of electric flux given by Eq. (1.28) can be used, in principle, to calculate the total flux through any given surface. All we have to do is to divide the surface into small

area elements, calculate the flux at each element and add them up. Thus, the total flux ϕ through a surface S is

$$\phi \simeq \sum \mathbf{E} \cdot \Delta \mathbf{S} \quad (1.29)$$

The approximation sign is put because the electric field \mathbf{E} is taken to be constant over the small area element. This is mathematically exact only when you take the limit $\Delta S \rightarrow 0$ and the sum in Eq. (1.29) is written as an integral. We do not discuss that procedure in detail here.

1.13 GAUSS'S THEOREM

As a simple application of the notion of electric flux, let us consider the total flux through a sphere, which encloses a point charge q at its centre.

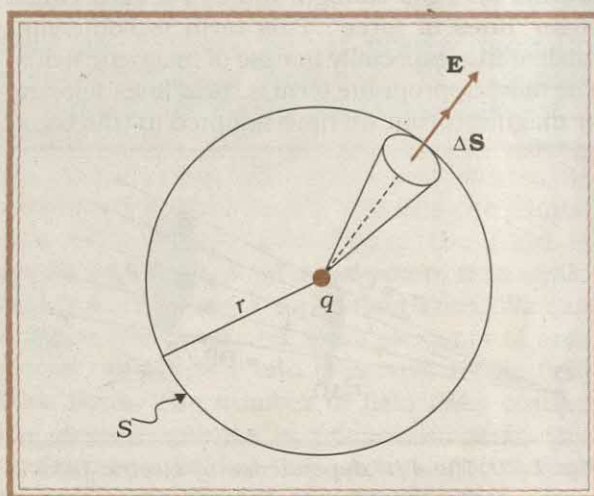


Fig. 1.23 Flux through a sphere enclosing a point charge q at its centre.

Divide the sphere into small area elements. The flux through an area element $\Delta \mathbf{S}$ is:

$$\Delta \phi = \mathbf{E} \cdot \Delta \mathbf{S} = \frac{q}{4 \pi \epsilon_0 r^2} \hat{\mathbf{r}} \cdot \Delta \mathbf{S} \quad (1.30)$$

where we have used Coulomb's law for the electric field due to a single charge q . The unit vector $\hat{\mathbf{r}}$ is along the radius vector from the centre to the area element. Now since the normal to a sphere at every point is along the radius vector at that point, the area element $\Delta \mathbf{S}$ and $\hat{\mathbf{r}}$ have the same direction. Therefore,

$$\Delta \phi = \frac{q}{4 \pi \epsilon_0 r^2} \Delta S \quad (1.31)$$

since the magnitude of $\hat{\mathbf{r}}$ is 1.

The total flux through the sphere is obtained by adding up flux through all the different area elements:

$$\phi = \sum_{\text{all } \Delta S} \frac{q}{4\pi\epsilon_0 r^2} \Delta S \quad (1.32)$$

Since each area element of the sphere is at the same distance r from the charge,

$$\phi = \frac{q}{4\pi\epsilon_0 r^2} \sum_{\text{all } \Delta S} \Delta S = \frac{q}{4\pi\epsilon_0 r^2} S \quad (1.33)$$

Now S , the total area of the sphere, equals $4\pi r^2$. Thus,

$$\phi = \frac{q}{4\pi\epsilon_0 r^2} \times 4\pi r^2 = \frac{q}{\epsilon_0} \quad (1.34)$$

Eq. (1.34) is a simple illustration of a general result of electrostatics called Gauss's theorem. We state Gauss's theorem without proof:

Electric flux through a closed surface S

$$= q/\epsilon_0, \quad (1.35)$$

q = total charge enclosed by S

The theorem implies that the total electric flux through a closed surface is zero if no charge is enclosed by the surface. We can see that explicitly in the simple situation of Fig. 1.24. Here the electric field is uniform and we are considering a closed cylindrical surface, with the axis of the cylinder parallel to the uniform field \mathbf{E} . The total flux ϕ through the surface is:

$$\phi = \phi_1 + \phi_2 + \phi_3 \quad (1.36)$$

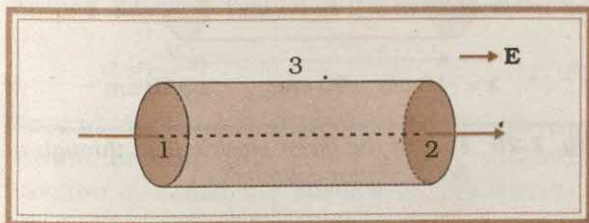


Fig. 1.24 Calculation of the flux of uniform electric field through the surface of a cylinder.

where ϕ_1 and ϕ_2 represent the flux through the surfaces 1 and 2 (of circular cross-section) of the cylinder and ϕ_3 is the flux through the curved cylindrical part of the closed surface. Now the normal to the surface 3 at every point is perpendicular to \mathbf{E} , so by definition of flux, $\phi_3 = 0$. Further, the outward normal to 2 is along \mathbf{E} while the outward normal to 1 is opposite to \mathbf{E} .

Therefore,

$$\begin{aligned} \phi_1 &= -E S_1, & \phi_2 &= +E S_2, \\ S_1 &= S_2 = S, \end{aligned} \quad (1.37)$$

where S is the area of circular cross-section. Thus, the total flux is zero, as expected by Gauss's theorem.

The great significance of Gauss's theorem Eq. (1.35), is that it is true in general, and not only for the simple cases we have considered above. Let us note some important points regarding this theorem:

- Gauss's theorem is true for any closed surface, no matter what its shape or size.
- The term q on the right side of Gauss's theorem, Eq. (1.35), includes the sum of all charges enclosed by the surface. The charges may be located anywhere inside the surface.
- In the situation when the surface is so chosen that there are some charges inside and some outside, the electric field [whose flux appears on the left side of Eq. (1.35)] is due to all the charges, both inside and outside S . The term q on the right side of Gauss's theorem, however, represents only the total charge inside S .
- The surface that we choose for the application of Gauss's law is called the Gaussian surface. You may choose any Gaussian surface and apply Gauss's theorem. However, take care not to let the Gaussian surface pass through any discrete charge. This is because electric field due to a system of discrete charges is not well defined at the location of any charge. (As you go close to the charge, the field grows without any bound.) However, the Gaussian surface can pass through a continuous charge distribution. (Think why?)
- For charge configurations with some symmetry, the choice of a suitable Gaussian surface is important, as it facilitates the calculation of electric field. In fact, as we shall see, Gauss's theorem is most commonly used for symmetric charge configurations.
- Finally, Gauss's theorem is based on the inverse square dependence on distance contained in the Coulomb's law. Any violation of Gauss's theorem will indicate departure from the inverse square law.

Example 1.8 The electric field components in the following figure are $E_x = \alpha x^{1/2}$, $E_y = E_z = 0$, in which $\alpha = 800 \text{ N/C m}^{1/2}$. Calculate (a) the flux ϕ_E through the cube, and (b) the charge within the cube. Assume that $a = 0.1 \text{ m}$.

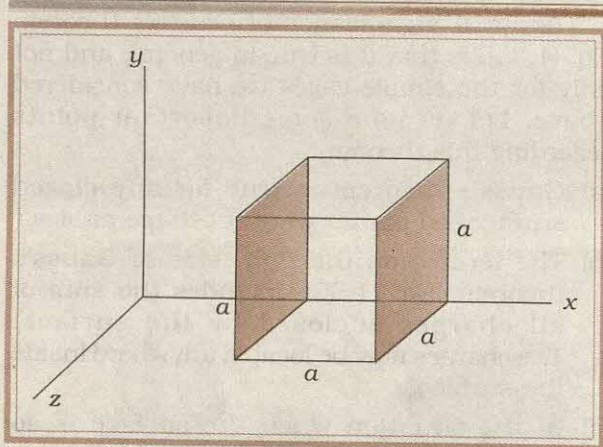


Fig. 1.25 Flux of the given electric field through a cube.

Answer

(a) Since the electric field has only an x component, for faces perpendicular to x direction, the angle between \mathbf{E} and $\Delta\mathbf{S}$ is $\pm \pi/2$. Therefore, the flux $\phi = \mathbf{E} \cdot \Delta\mathbf{S}$ is separately zero for each face of the cube except the two shaded ones. Now the magnitude of the electric field at the left face is

$$E_L = \alpha x^{1/2} = \alpha a^{1/2}$$

($x = a$ at the left face).

The magnitude of electric field at the right face is

$$E_R = \alpha x^{1/2} = \alpha (2a)^{1/2}$$

($x = 2a$ at the right face).

The corresponding fluxes are

$$\begin{aligned}\phi_L &= \mathbf{E}_L \cdot \Delta\mathbf{S} = E_L \Delta S \cos \theta \\ &= -E_L \Delta S, \text{ since } \theta = 180^\circ \\ &= -E_L a^2\end{aligned}$$

$$\begin{aligned}\phi_R &= \mathbf{E}_R \cdot \Delta\mathbf{S} = E_R \Delta S \cos \theta \\ &= E_R \Delta S, \text{ since } \theta = 0^\circ \\ &= E_R a^2\end{aligned}$$

Net flux through the cube

$$\begin{aligned}&= \phi_L + \phi_R = E_R a^2 - E_L a^2 \\ &= a^2 (E_R - E_L) \\ &= \alpha a^2 [(2a)^{1/2} - a^{1/2}] \\ &= \alpha a^{5/2} (\sqrt{2} - 1)\end{aligned}$$

$$= 800 (0.1)^{5/2} (\sqrt{2} - 1)$$

$$= 1.05 \text{ N m}^2 \text{ C}^{-1}$$

(b) We can use Gauss's theorem to find the total charge q inside the cube. We have $\phi = q/\epsilon_0$ or $q = \phi \epsilon_0$. Therefore,

$$q = 1.05 \times 8.854 \times 10^{-12} \text{ C}$$

$$= 9.27 \times 10^{-12} \text{ C}.$$

Example 1.9 An electric field is uniform, and in the positive x direction for positive x , and uniform with the same magnitude but in the negative x direction for negative x . It is given that $\mathbf{E} = 200 \hat{\mathbf{i}} \text{ N/C}$ for $x > 0$ and $\mathbf{E} = -200 \hat{\mathbf{i}} \text{ N/C}$ for $x < 0$. A right circular cylinder of length 20 cm and radius 5 cm has its centre at the origin and its axis along the x -axis so that one face is at $x = +10 \text{ cm}$ and the other is at $x = -10 \text{ cm}$. (a) What is the net outward flux through each flat face? (b) What is the flux through the side of the cylinder? (c) What is the net outward flux through the cylinder? (d) What is the net charge inside the cylinder?

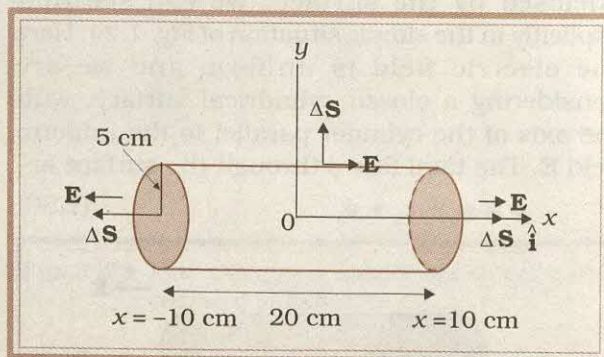


Fig. 1.26 Flux of the given electric field through a right circular cylinder.

Answer

(a) We can see from the figure that on the left face \mathbf{E} and $\Delta\mathbf{S}$ are parallel. Therefore, the outward flux is

$$\begin{aligned}\phi_L &= \mathbf{E} \cdot \Delta\mathbf{S} = -200 \hat{\mathbf{i}} \cdot \Delta\mathbf{S} \\ &= +200 \Delta S, \text{ since } \hat{\mathbf{i}} \cdot \Delta\mathbf{S} = -\Delta S \\ &= +200 \times \pi (0.05)^2 = +1.57 \text{ N m}^2 \text{ C}^{-1}.\end{aligned}$$

On the right face, \mathbf{E} and $\Delta\mathbf{S}$ are parallel and therefore

$$\begin{aligned}\phi_R &= \mathbf{E} \cdot \Delta\mathbf{S} \\ &= +1.57 \text{ N m}^2 \text{ C}^{-1}.\end{aligned}$$

(b) For any point on the side of the cylinder \mathbf{E} is perpendicular to $\Delta\mathbf{S}$ and hence $\mathbf{E} \cdot \Delta\mathbf{S} = 0$. Therefore, the flux out of the side of the cylinder is zero.

(c) Net outward flux through the cylinder
 $\phi = 1.57 + 1.57 + 0 = 3.14 \text{ N m}^2 \text{ C}^{-1}$.

(d) The net charge within the cylinder can be found by using Gauss's theorem which gives

$$\begin{aligned} q &= \epsilon_0 \phi \\ &= 3.14 \times 8.854 \times 10^{-12} \text{ C} \\ &= 2.78 \times 10^{-11} \text{ C}. \end{aligned}$$

1.14 CONTINUOUS CHARGE DISTRIBUTION

We have so far dealt with charge configurations involving discrete charges q_1, q_2, \dots, q_n . One reason why we restricted to discrete charges is that the mathematical treatment is simpler and does not involve calculus. For many purposes, however, it is impractical to work in terms of discrete charges and we need to work with continuous charge distributions. For example, on the surface of a charged conductor, it is impractical to specify the charge distribution in terms of the locations of the microscopic charged constituents. It is more feasible to consider an area element ΔS (Fig. 1.27) on the surface of the conductor (which is very small on the macroscopic scale but big enough to include a very large number of electrons) and specify the charge ΔQ on that element. We then define a surface charge density σ at the area element by

$$\sigma = \left(\frac{\Delta Q}{\Delta S} \right) \quad (1.38)$$

We can do this at different points on the conductor and thus arrive at a continuous function σ , called the surface charge density. The surface charge density σ so defined ignores the quantisation of charge and the discontinuity in charge distribution at the microscopic level*. σ represents macroscopic surface charge density, which in a sense, is a smoothed out average of the microscopic charge density over an area element ΔS which, as said before, is large microscopically but small macroscopically.

Similar considerations apply for a line charge distribution and a volume charge distribution.

The linear charge density λ of a wire is defined by

$$\lambda = \frac{\Delta Q}{\Delta l} \quad (1.39)$$

where Δl is a small line element of wire on the macroscopic scale that, however, includes a large number of microscopic charged constituents, and ΔQ is the charge contained in that line element. The volume charge density (sometimes simply called charge density) is defined in a similar manner:

$$\rho = \frac{\Delta Q}{\Delta V} \quad (1.40)$$

where ΔQ is the charge included in the macroscopically small volume element ΔV that includes a large number of microscopic charged constituents.

The notion of continuous charge distribution is similar to that we adopt for continuous mass distribution in mechanics. When we refer to the density of a liquid, we are referring to its macroscopic density. We regard it as a continuous fluid and ignore its discrete molecular constitution.

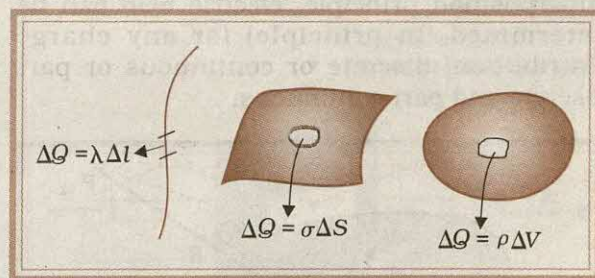


Fig. 1.27 Definitions of linear, surface and volume charge densities. In each case, the element ($\Delta l, \Delta S, \Delta V$) chosen is small on the macroscopic scale but contains a very large number of microscopic constituents.

The field due to a continuous charge distribution can be obtained in much the same way as for a system of discrete charges, Eq. (1.16). Suppose a continuous charge distribution in space has a charge density ρ . Choose any convenient origin O and let the position vector of any point in the charge distribution be \mathbf{r} . The charge density ρ may vary from point to point i.e., it is a function of \mathbf{r} . Divide the charge distribution into small volume

* At the microscopic level, charge distribution is discontinuous, because there are discrete charges separated by intervening space where there is no charge.

elements of size ΔV . The charge in a volume element ΔV is $\rho \Delta V$.

Consider any general point P (inside or outside the distribution) with position vector \mathbf{R} (Fig. 1.28). Electric field due to the charge $\rho \Delta V$ is given by Coulomb's Law:

$$\Delta \mathbf{E} = \frac{1}{4\pi\epsilon_0} \frac{\rho \Delta V}{r'^2} \hat{\mathbf{r}}' \quad (1.41)$$

where r' is the distance between the charge element and P, and $\hat{\mathbf{r}}'$ is a unit vector in the direction from the charge element to P. By the superposition principle, the total electric field due to the charge distribution is given by summing over electric fields due to different volume elements:

$$\mathbf{E} \equiv \frac{1}{4\pi\epsilon_0} \sum_{\text{all } \Delta V} \frac{\rho \Delta V}{r'^2} \hat{\mathbf{r}}' \quad (1.42)$$

Note that ρ , r' , $\hat{\mathbf{r}}'$ all vary from point to point. In a strict mathematical method, we should let $\Delta V \rightarrow 0$ and the sum then becomes an integral; but we omit that discussion here, for simplicity. In short, using Coulomb's law and the superposition principle, electric field can be determined (in principle) for any charge distribution, discrete or continuous or part discrete and part continuous.

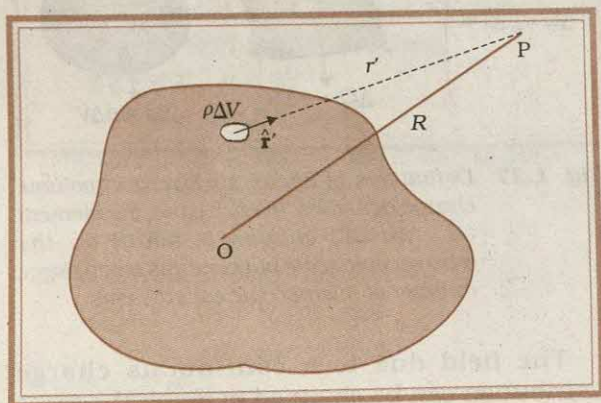


Fig. 1.28 Notations used in Eq. (1.42) for field due to a continuous charge distribution.

1.15 APPLICATIONS OF GAUSS'S THEOREM

The electric field due to a general charge distribution is, as seen above, given by Eq. (1.42). In practice, except for some special cases, the summation (or integration) involved in this equation cannot be carried out to give electric field at every point in space.

For some symmetric charge configurations, however, it is possible to obtain the electric field in a simple way using the Gauss's theorem. This is best understood by some examples.

Field due to an infinitely long straight charged wire

Consider an infinitely long thin straight wire with uniform linear charge density λ . The wire is obviously an axis of symmetry. Suppose we take the radial vector from O to P and rotate it around the wire. The points P, P', P'' so obtained are completely equivalent with respect to the charged wire. This implies that the electric field must have the same magnitude at these points. The direction of electric field at every point must be radial (outward if $\lambda > 0$, inward if $\lambda < 0$). This is clear from Fig. 1.29. Consider a pair of line elements of the wire, as shown. The electric fields produced by the two elements of the pair when summed give a resultant electric field which is radial (the components normal to the radial vector cancel). This is true for any pair and hence the total field at any point P is radial. Finally, since the wire is infinite, electric field does not depend on the position of P along the length of the wire. In short, the electric field is everywhere radial in the plane cutting the wire normally, and its magnitude depends only on the radial distance r .

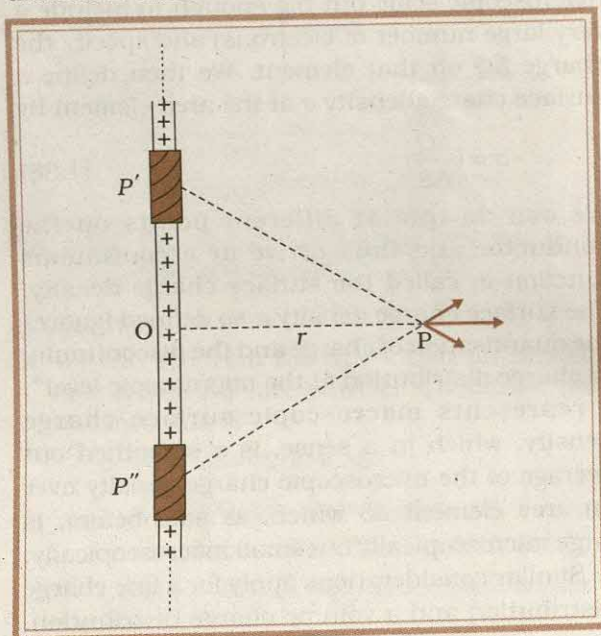


Fig. 1.29 Electric field due to an infinitely long thin straight wire is radial.

To calculate the field, imagine a cylindrical Gaussian surface, as shown in Fig. 1.30.

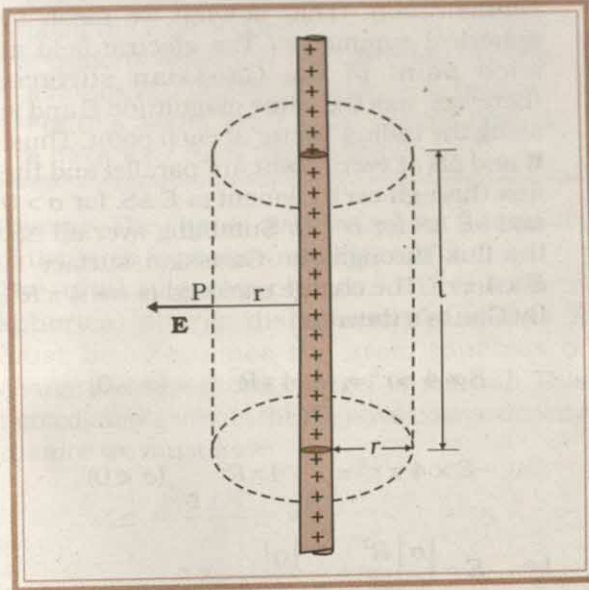


Fig. 1.30 The Gaussian surface for a long thin wire of uniform linear charge density.

Since the field is everywhere radial, flux through the two ends of the cylindrical Gaussian surface is zero. At the cylindrical part of the surface, \mathbf{E} is normal to the surface at every point, and its magnitude is constant, since it depends only on r . The surface area of the curved part is $2\pi r l$, where l is the length of the cylinder. Flux through the Gaussian surface

= flux through the curved cylindrical part of the surface

$$= E \times 2\pi r l \quad (\text{for } \lambda > 0)$$

$$= -E \times 2\pi r l \quad (\text{for } \lambda < 0)$$

since for $\lambda < 0$, \mathbf{E} is inward.

The surface includes charge equal to λl . Gauss's theorem then gives

$$E \times 2\pi r l = \frac{\lambda l}{\epsilon_0} \quad (\lambda > 0)$$

$$-E \times 2\pi r l = \frac{\lambda l}{\epsilon_0} \quad (\lambda < 0)$$

$$\text{i.e., } E = \frac{|\lambda|}{2\pi\epsilon_0 r}$$

Vectorially, for either sign of λ , \mathbf{E} at any point is given by

$$\mathbf{E} = \frac{\lambda}{2\pi\epsilon_0 r} \hat{\mathbf{n}} \quad (1.43)$$

where $\hat{\mathbf{n}}$ is the radial unit vector in the plane normal to the wire passing through the point.

Note that though only the charge enclosed by the surface (λl) was included above, the electric field \mathbf{E} is due to the charge on the entire wire. Further, the assumption that the wire is infinitely long is crucial. Without this assumption, we cannot take \mathbf{E} to be normal to the curved part of the cylindrical Gaussian surface. However, Eq. (1.43) is approximately true for electric field around the central portions of a long wire, where the end effects may be ignored.

Field due to a uniformly charged infinite plane sheet

Let σ be the uniform surface charge density of an infinite plane sheet (Fig. 1.31). We take the x -axis normal to the given plane. By symmetry, the electric field will not depend on y and z co-ordinates and its direction at every point must be along the x direction.

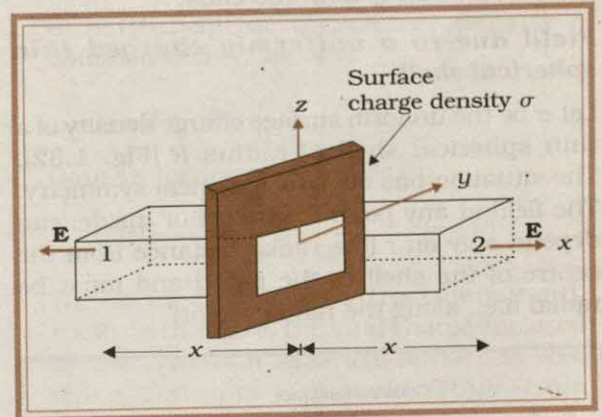


Fig. 1.31 Gaussian surface for a uniformly charged infinite plane sheet.

We can take the Gaussian surface to be a rectangular parallelepiped of cross sectional area A , as shown. (A cylindrical surface will also do.) As seen from the figure, only the two faces 1 and 2 will contribute to the flux; electric field lines are parallel to the other faces and they, therefore, do not contribute to the total flux.

Flux through the Gaussian surface

= flux through the two faces normal to \mathbf{E}

$$= 2 E \times A \quad \text{if } \sigma > 0$$

$$= -2 E \times A \quad \text{if } \sigma < 0$$

The charge enclosed by the closed surface $= \sigma \times A$. By Gauss's theorem,

$$2EA = \frac{\sigma A}{\epsilon_0} \quad (\sigma > 0)$$

$$-2EA = \frac{\sigma A}{\epsilon_0} \quad (\sigma < 0)$$

$$\text{i.e., } E = \frac{|\sigma|}{2\epsilon_0}$$

Vectorially, for either sign of σ ,

$$\mathbf{E} = \frac{\sigma}{2\epsilon_0} \hat{\mathbf{n}} \quad (1.44)$$

where $\hat{\mathbf{n}}$ is a unit vector normal to the plane and going away from it. If $\sigma > 0$, the electric field is uniform, normal and outward from the sheet. For $\sigma < 0$, the direction of \mathbf{E} is along the inward normal to the plane.

For a finite large planar sheet, Eq. (1.44) is approximately true in the middle regions of the planar sheet away from the ends.

Field due to a uniformly charged thin spherical shell

Let σ be the uniform surface charge density of a thin spherical shell of radius R (Fig. 1.32). The situation has obvious spherical symmetry. The field at any point P, outside or inside, can depend only on r (the radial distance from the centre of the shell to the point) and must be radial (i.e., along the radial vector).

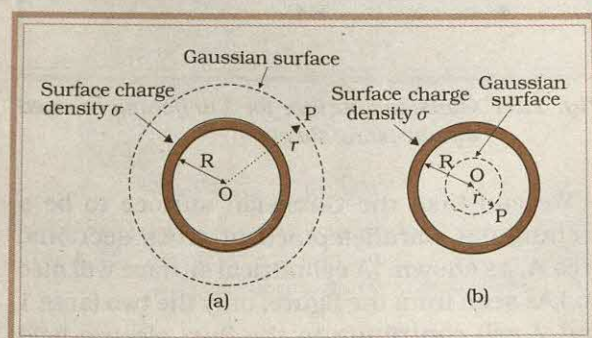


Fig. 1.32 The Gaussian surface for the case of a thin uniformly charged spherical shell.

- (a) *Field outside the shell:* Consider a point P outside the shell with radial vector \mathbf{r} . To calculate \mathbf{E} at P, we take the Gaussian surface to be a sphere of radius r and with centre O,

passing through P. All points on this sphere are equivalent relative to the given charged configuration. (That is what we mean by spherical symmetry.) The electric field at each point of the Gaussian surface, therefore, has the same magnitude E and is along the radius vector at each point. Thus, \mathbf{E} and $\Delta\mathbf{S}$ at every point are parallel and the flux through each element is $E \Delta S$, for $\sigma > 0$ and $-E \Delta S$ for $\sigma < 0$. Summing over all ΔS , the flux through the Gaussian surface is $E \times 4\pi r^2$. The charge enclosed is $\sigma \times 4\pi R^2$. By Gauss's theorem

$$E \times 4\pi r^2 = \frac{\sigma}{\epsilon_0} 4\pi R^2 \quad (\sigma > 0)$$

$$-E \times 4\pi r^2 = \frac{\sigma}{\epsilon_0} 4\pi R^2 \quad (\sigma < 0)$$

$$\text{i.e., } E = \frac{|\sigma| R^2}{\epsilon_0 r^2} = \frac{|q|}{4\pi\epsilon_0 r^2}$$

where $q = 4\pi R^2 \sigma$ is the total charge on the spherical shell.

Vectorially,

$$\mathbf{E} = \frac{q}{4\pi\epsilon_0 r^2} \hat{\mathbf{r}} \quad (1.45)$$

This, however, is exactly the field produced by a charge q placed at the centre O. Thus for points outside the shell, the field due to a uniformly charged shell is as if the entire charge of the shell is concentrated at its centre.

- (b) *Field inside the shell:* In Fig. 1.32(b), the point P is inside the shell. The Gaussian surface is again a sphere through P centred at O. The flux through the Gaussian surface, calculated as before, is $E \times 4\pi r^2$ (to within a sign). However, in this case, the Gaussian surface encloses no charge. Gauss's theorem then gives

$$E \times 4\pi r^2 = 0$$

$$\text{i.e., } E = 0 \quad (r < R) \quad (1.46)$$

i.e., the field due to a uniformly charged thin shell is zero at all points inside the shell. This important result is a direct consequence of Gauss's law which follows from Coulomb's law. The experimental verification of this result confirms the $1/r^2$ dependence in Coulomb's law.

Example 1.10 An early model for an atom considered it to have a positively charged point nucleus of charge Ze , surrounded by a uniform density of negative charge up to a radius R . The atom as a whole is neutral. For this model, what is the electric field at a distance r from the nucleus?

Answer The charge distribution for this model of the atom is as shown in Fig. 1.33. The total negative charge in the uniform spherical charge distribution of radius R must be $-Ze$, since the atom (nucleus of charge Ze + negative charge) is neutral. This immediately gives us the negative charge density ρ , since we must have

$$Ze + \frac{4\pi R^3}{3} \rho = 0$$

or
$$\rho = -\frac{3Ze}{4\pi R^3}$$

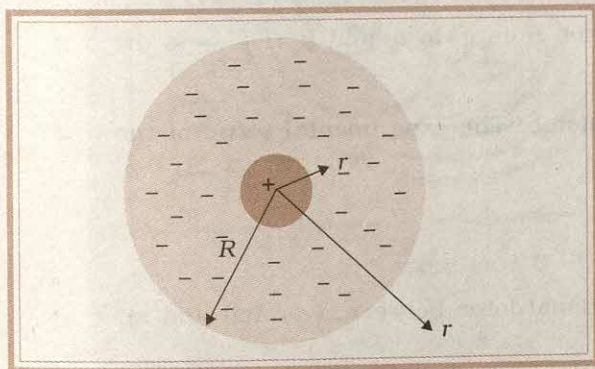


Fig. 1.33 Early model of an atom.

To find the electric field $\mathbf{E}(\mathbf{r})$ at a point P which is a distance r away from the nucleus, we use Gauss's theorem. Because of the spherical symmetry of the charge distribution, the

magnitude of the electric field $\mathbf{E}(\mathbf{r})$ depends only on the radial distance, no matter what the direction of \mathbf{r} . Its direction is along (or opposite to) the radius vector \mathbf{r} from the origin to the point P. The obvious Gaussian surface is a spherical surface centred at the nucleus. We consider two situations, namely, $r < R$ and $r > R$.

(a) $r < R$: The electric flux ϕ enclosed by the spherical surface is

$$\phi = E(r) \times 4\pi r^2$$

where $E(r)$ is the magnitude of the electric field at r . This is because the field at any point on the spherical Gaussian surface has the same direction as the normal to the surface there, and has the same magnitude at all points on the surface.

The charge q enclosed by the Gaussian surface is the positive nuclear charge and the negative charge within the sphere of radius r ,

i.e.,
$$q = Ze + \frac{4\pi r^3}{3} \rho$$

Substituting for the charge density ρ obtained earlier, we have

$$q = Ze - Ze \frac{r^3}{R^3}$$

Gauss's theorem then gives,

$$E(r) = \frac{Ze}{4\pi\epsilon_0} \left(\frac{1}{r^2} - \frac{r}{R^3} \right); \quad (r < R)$$

The electric field is directed radially outward.

(b) $r > R$: In this case, the total charge enclosed by the Gaussian spherical surface is zero since the atom is neutral. Thus, from Gauss's theorem,

$$E(r) 4\pi r^2 = 0 \text{ or } E(r) = 0; \quad (r > R)$$

At $r = R$, both cases give the same result: $E = 0$.

SUMMARY

1. Electric and magnetic forces determine the properties of atoms, molecules and bulk matter.
2. From simple experiments on frictional electricity, one can infer that there are two types of charge in nature; and that like charges repel and unlike charges attract. By convention, the charge on a glass rod rubbed with silk is positive; that on a plastic rod rubbed with fur is then negative.

3. Conductors allow large scale movement of electric charge through them, insulators do not. In metals, the mobile charges are electrons; in electrolytes both positive and negative ions are mobile.
4. Electric charge has three basic properties: quantisation, additivity and conservation.

Quantisation of electric charge means that total charge (q) of a body is always an integral multiple of a basic quantum of charge (e) i.e., $q = n e$, where $n = 0, \pm 1, \pm 2, \pm 3, \dots$. Proton and electron have charges $+e, -e$, respectively. For macroscopic charges for which n is a very large number, quantisation of charge can be ignored.

Additivity of electric charge means that the total charge of a system is the algebraic sum (i.e., the sum taking into account proper signs) of all individual charges in the system.

Conservation of electric charge means that the total charge of an isolated system remains unchanged with time. This means that when bodies are charged through friction, there is a transfer of electric charge from one body to another, but no creation or destruction of charge.

5. **Coulomb's Law:** The mutual electrostatic force between two point charges q_1 and q_2 is proportional to the product $q_1 q_2$ and inversely proportional to the square of the distance r_{21} separating them. Mathematically,

$$\mathbf{F}_{21} = \text{force on } q_2 \text{ due to } q_1 = \frac{k(q_1 q_2)}{r_{21}^2} \hat{\mathbf{r}}_{21}$$

where $\hat{\mathbf{r}}_{21}$ is a unit vector in the direction from q_1 to q_2 and $k = \frac{1}{4\pi\epsilon_0}$ is the

constant of proportionality.

In SI units, the unit of charge is coulomb. The experimental value of the constant ϵ_0 is

$$\epsilon_0 = 8.854 \times 10^{-12} \text{ C}^2 \text{ N}^{-1} \text{ m}^{-2}$$

The approximate value of k is

$$k = 9 \times 10^9 \text{ N m}^2 \text{ C}^{-2}$$

6. The ratio of electric force and gravitational force between a proton and an electron is

$$\frac{ke^2}{Gm_e m_p} \approx 2.4 \times 10^{39}$$

7. **Superposition Principle:** The principle is based on the property that the forces with which two charges attract or repel each other are not affected by the presence of a third (or more) additional charge(s). For an assembly of charges q_1, q_2, q_3, \dots , the force on any charge, say q_1 , is the vector sum of the force on q_1 due to q_2 , the force on q_1 due to q_3 , and so on. For each pair, the force is given by the Coulomb's law for two charges stated earlier.
8. The electric field \mathbf{E} at a point due to a charge configuration is the force on a small positive test charge q placed at the point divided by the magnitude of the charge. Electric field due to a point charge q has a magnitude $|q|/4\pi\epsilon_0 r^2$; it is radially outwards from q , if q is positive, and radially inwards if q is negative. Like Coulomb force, electric field also satisfies superposition principle.
9. An electric field line is a curve drawn in such a way that the tangent at each point on the curve gives the direction of electric field at that point. The relative closeness of field lines indicates the relative strength of electric field at different

points; they crowd near each other in regions of strong electric field and are far apart where the electric field is weak. In regions of constant electric field, the field lines are parallel straight lines.

10. Some of the important properties of field lines are: (a) Field lines are continuous curves without any breaks. (b) Two field lines cannot cross each other. (c) Electrostatic field lines start at positive charges and end at negative charges—they cannot form closed loops.
11. An electric dipole is a pair of equal and opposite charges q and $-q$ separated by some distance $2a$. Its dipole moment vector \mathbf{p} has magnitude $2qa$ and is in the direction of the dipole axis from $-q$ to q .
12. Field of an electric dipole in its equatorial plane (i.e., the plane perpendicular to its axis and passing through its centre) at a distance r from the centre:

$$\mathbf{E} = \frac{-\mathbf{p}}{4\pi\epsilon_0 (a^2 + r^2)^{3/2}}$$

$$\approx \frac{-\mathbf{p}}{4\pi\epsilon_0 r^3}, \quad \text{for } r \gg a$$

Dipole electric field on the axis at a distance r from the centre :

$$\mathbf{E} = \frac{2\mathbf{p}r}{4\pi\epsilon_0 (r^2 - a^2)^2}$$

$$\approx \frac{2\mathbf{p}}{4\pi\epsilon_0 r^3}, \quad \text{for } r \gg a$$

The $1/r^3$ dependence of dipole electric fields should be noted in contrast to the $1/r^2$ dependence of electric field due to a point charge.

13. In a uniform electric field \mathbf{E} , a dipole experiences a torque τ given by

$$\tau = \mathbf{p} \times \mathbf{E}$$

but experiences no net force.

14. The flux $\Delta\phi$ of electric field \mathbf{E} through a small area element $\Delta\mathbf{S}$ is given by

$$\Delta\phi = \mathbf{E} \cdot \Delta\mathbf{S}$$

The vector area element $\Delta\mathbf{S}$ is

$$\Delta\mathbf{S} = \Delta S \hat{\mathbf{n}}$$

where ΔS is the magnitude of the area element and $\hat{\mathbf{n}}$ is normal to the area element, which can be considered planar for sufficiently small ΔS . For an area element of a closed surface, $\hat{\mathbf{n}}$ is taken to be the direction of outward normal, by convention.

15. **Gauss's Theorem:** The flux of electric field through any closed surface S is $1/\epsilon_0$ times the total charge enclosed by S . The theorem is especially useful in determining electric field \mathbf{E} , when the source distribution has simple symmetry:
 - (i) Thin infinitely long straight wire of uniform linear charge density λ

$$\mathbf{E} = \frac{\lambda}{2\pi\epsilon_0 r} \hat{\mathbf{n}}$$

where r is the perpendicular distance of the point from the wire and $\hat{\mathbf{n}}$ is the radial unit vector in the plane normal to the wire passing through the point.

- (ii) Infinite thin plane sheet of uniform surface charge density σ

$$\mathbf{E} = \frac{\sigma}{2\epsilon_0} \hat{\mathbf{n}}$$

where \hat{n} is a unit vector normal to the plane, outward on either side.
 (iii) *Thin spherical shell of uniform surface charge density σ*

$$\mathbf{E} = \frac{q}{4\pi\epsilon_0 r^2} \hat{\mathbf{r}} \quad (r \geq R)$$

$$\mathbf{E} = 0 \quad (r < R)$$

where r is the distance of the point from the centre of the shell and R the radius of the shell. q is the total charge of the shell: $q = 4\pi R^2 \sigma$.
 The electric field outside the shell is as though the total charge is concentrated at the centre. The same result is true for a solid sphere of uniform volume charge density.

Physical quantity	Symbol	Dimensions	Unit	Remarks
Vector area element	$\Delta \mathbf{S}$	$[L^2]$	m^2	$\Delta \mathbf{S} = \Delta S \hat{\mathbf{n}}$
Electric field	\mathbf{E}	$[MLT^{-3}A^{-1}]$	$V m^{-1}$	
Electric flux	ϕ	$[ML^3 T^{-3} A^{-1}]$	$V m$	$\Delta \phi = \mathbf{E} \cdot \Delta \mathbf{S}$
Dipole moment	\mathbf{p}	$[LTA]$	$C m$	Vector directed from negative to positive charge
Charge density				
linear	λ	$[L^{-1} TA]$	$C m^{-1}$	Charge/length
surface	σ	$[L^{-2} TA]$	$C m^{-2}$	Charge/area
volume	ρ	$[L^{-3} TA]$	$C m^{-3}$	Charge/volume

POINTS TO PONDER

1. The constant of proportionality k in Coulomb's law is a matter of choice if the unit of charge is to be defined using Coulomb's law. In SI units, however, what is defined is the unit of current (A) via its magnetic effect (Ampere's law) and the unit of charge (coulomb) is simply defined by (1C = 1 A s). In this case, the value of k is no longer arbitrary; it is approximately $9 \times 10^9 N m^2 C^{-2}$.
2. The rather large value of k , i.e., the large size of the unit of charge (1C) from the point of view of electric effects arises because (as mentioned in point 1 already) the unit of charge is defined in terms of magnetic forces (forces on current-carrying wires) which are generally much weaker than the electric forces. This is why, while 1 ampere is a unit of reasonable size for magnetic effects, 1 C = 1 A s, is too big a unit for electric effects.
3. The additive property of charge is not an 'obvious' property. It is related to the fact that electric charge has no direction associated with it; charge is a scalar.
4. Charge is not only a scalar (or invariant) under rotation; it is also invariant for frames of reference in relative motion. This is not always true. For example, kinetic energy is a scalar under rotation, but is not invariant for frames of reference in relative motion.
5. Conservation of total charge of an isolated system is a property independent of the scalar nature of charge noted in point 4. Conservation refers to invariance in time in a given frame of reference. A quantity may be scalar but not conserved (like kinetic energy in an inelastic collision). On the other hand, one can have conserved vector quantity (e.g. angular momentum of an isolated system).

6. Quantisation of electric charge is a basic (unexplained) law of nature; interestingly, there is no analogous law on quantisation of mass.
7. Superposition principle should not be regarded as 'obvious', or equated with the law of addition of vectors. It says two things: force on one charge due to another charge is unaffected by the presence of other charges, and there are no additional three-body, four-body, etc., forces which arise only when there are more than two charges.
8. The electric field due to a discrete charge configuration is not defined at the location of the discrete charges. For continuous volume charge distribution, it is defined at any point in the distribution. For a surface charge distribution, electric field is discontinuous across the surface.
9. The electric field due to a charge configuration with total charge zero is not zero; but for distances large compared to the size of the configuration, its field falls off faster than $1/r^2$, typical of field due to a single charge. An electric dipole is the simplest example of this fact.
10. Coulomb force and gravitational force follow the same inverse-square law. But gravitational force has only one sign (always attractive), while Coulomb force can be of both signs (attractive and repulsive), allowing possibility of cancellation of electric forces. This is how gravity, despite being a much weaker force, can be a dominating and more pervasive force in nature.

EXERCISES

- 1.1 State Coulomb's law of force between two charges at rest. What is the force of repulsion between two charges of 1C each, kept 1m apart in vacuum?
- 1.2 What is the force between two small charged spheres having charges of $2 \times 10^{-7}\text{C}$ and $3 \times 10^{-7}\text{C}$ placed 30 cm apart in air?
- 1.3 The electrostatic force on a small sphere of charge $0.4 \mu\text{C}$ due to another small sphere of charge $-0.8 \mu\text{C}$ in air is 0.2 N. (a) What is the distance between the two spheres? (b) What is the force on the second sphere due to the first?
- 1.4 Check that the ratio ke^2/Gm_em_p is dimensionless. Look up a Table of Physical Constants and determine the value of this ratio. What does the ratio signify?
- 1.5 (a) Explain the meaning of the statement 'electric charge of a body is quantised'.
(b) Why can one ignore quantisation of electric charge when dealing with macroscopic i.e., large scale charges?
- 1.6 When a glass rod is rubbed with a silk cloth, charges appear on both. A similar phenomenon is observed with many other pairs of bodies. Explain how this observation is consistent with the law of conservation of charge.
- 1.7 State the superposition principle for electrostatic force on a charge due to a number of charges.
- 1.8 Four point charges $q_A = 2 \mu\text{C}$, $q_B = -5 \mu\text{C}$, $q_C = 2 \mu\text{C}$, and $q_D = -5 \mu\text{C}$ are located at the corners of a square ABCD of side 10 cm. What is the force on a charge of $1 \mu\text{C}$ placed at the centre of the square?

- 1.9** Define electric field \mathbf{E} at a point in space due to a distribution of charges. A point charge q is placed at the origin. How does the electric field due to the charge vary with distance r from the origin?
- 1.10** (a) An electrostatic field line is a continuous curve. That is, a field line cannot have sudden breaks. Why not?
 (b) Explain why two field lines never cross each other at any point?
- 1.11** Two point charges $q_A = 3 \mu\text{C}$ and $q_B = -3 \mu\text{C}$ are located 20 cm apart in vacuum.
 (a) What is the electric field at the midpoint O of the line AB joining the two charges?
 (b) If a negative test charge of magnitude $1.5 \times 10^{-9} \text{ C}$ is placed at this point, what is the force experienced by the test charge?
- 1.12** An electric dipole is placed in a uniform external electric field \mathbf{E} . Show that the torque on the dipole is given by

$$\boldsymbol{\tau} = \mathbf{p} \times \mathbf{E}$$
 where \mathbf{p} is the dipole moment of the dipole. What is the net force experienced by the dipole?
- 1.13** A system has two charges $q_A = 2.5 \times 10^{-7} \text{ C}$ and $q_B = -2.5 \times 10^{-7} \text{ C}$ located at points A: (0, 0, -15 cm) and B: (0, 0, +15 cm), respectively. What are the total charge and electric dipole moment of the system?
- 1.14** An electric dipole with dipole moment $4 \times 10^{-9} \text{ C m}$ is aligned at 30° with the direction of a uniform electric field of magnitude $5 \times 10^4 \text{ NC}^{-1}$. Calculate the magnitude of the torque acting on the dipole.
- 1.15** A polythene piece rubbed with wool is found to have a negative charge of $3 \times 10^{-7} \text{ C}$.
 (a) Estimate the number of electrons transferred (from which to which?)
 (b) Is there a transfer of mass from wool to polythene?
- 1.16** (a) Two insulated charged copper spheres A and B have their centres separated by a distance of 50 cm. What is the mutual force of electrostatic repulsion if the charge on each is $6.5 \times 10^{-7} \text{ C}$? The radii of A and B are negligible compared to the distance of separation.
 (b) What is the force of repulsion if each sphere is charged double the above amount, and the distance between them is halved?
- 1.17** Suppose the spheres A and B in Exercise 1.16 have identical sizes. A third sphere of the same size but uncharged is brought in contact with the first, then brought in contact with the second, and finally removed from both. What is the new force of repulsion between A and B?
- 1.18** Figure 1.34 shows tracks of three charged particles in a uniform electrostatic field. Give the signs of the three charges. Which particle has the highest charge to mass ratio?

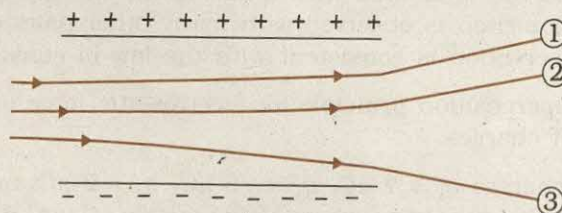


Fig. 1.34

- 1.19** Consider a uniform electric field $\mathbf{E} = 3 \times 10^3 \hat{\mathbf{i}} \text{ N/C}$. (a) What is the flux of this field through a square of 10 cm on a side whose plane is parallel to the yz plane? (b) What is the flux through the same square if the normal to its plane makes a 60° angle with the x -axis?
- 1.20** What is the net flux of the uniform electric field of Exercise 1.19 through a cube of side 20 cm oriented so that its faces are parallel to the coordinate planes?
- 1.21** Careful measurement of the electric field at the surface of a black box indicates that the net outward flux through the surface of the box is $8.0 \times 10^3 \text{ Nm}^2/\text{C}$. (a) What is the net charge inside the box? (b) If the net outward flux through the surface of the box were zero, could you conclude that there were no charges inside the box? Why or Why not?
- 1.22** A point charge $+10 \mu\text{C}$ is a distance 5 cm directly above the centre of a square of side 10 cm as shown in Fig. 1.35. What is the magnitude of the electric flux through the square? (Hint: Think of the square as one face of a cube with edge 10 cm.)

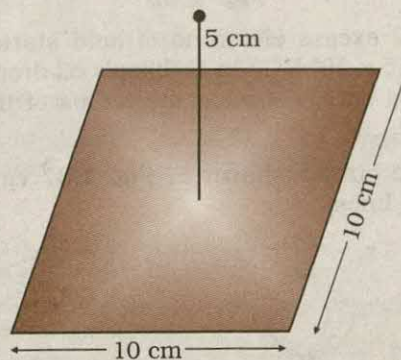


Fig. 1.35

- 1.23** A point charge of $2.0 \mu\text{C}$ is at the centre of a cubic Gaussian surface 9.0 cm on edge. What is the net electric flux through the surface?
- 1.24** A point charge causes an electric flux of $-1.0 \times 10^3 \text{ Nm}^2/\text{C}$ to pass through a spherical Gaussian surface of 10.0 cm radius centred on the charge. (a) If the radius of the Gaussian surface were doubled, how much flux would pass through the surface? (b) What is the value of the point charge?
- 1.25** A conducting sphere of radius 10 cm has an unknown charge. If the electric field 20 cm from the centre of the sphere is $1.5 \times 10^3 \text{ N/C}$ and points radially inward, what is the net charge on the sphere?
- 1.26** A uniformly charged conducting sphere of 2.4 m diameter has a surface charge density of $80.0 \mu\text{C}/\text{m}^2$. (a) Find the charge on the sphere. (b) What is the total electric flux leaving the surface of the sphere?
- 1.27** An infinite line charge produces a field of $9 \times 10^4 \text{ N/C}$ at a distance of 2 cm. Calculate the linear charge density.
- 1.28** Two large, thin metal plates are parallel and close to each other. On their inner faces, the plates have surface charge densities of opposite signs and of magnitude $17.0 \times 10^{-22} \text{ C}/\text{m}^2$. What is \mathbf{E} : (a) in the outer region of the first plate, (b) in the outer region of the second plate, and (c) between the plates?

ADDITIONAL EXERCISES

- 1.29** A glass rod rubbed with silk is brought close to two uncharged metallic spheres in contact with each other, inducing charges on them as shown in Fig. 1.36. Describe what happens when
- the spheres are slightly separated, and
 - the glass rod is subsequently removed, and finally
 - the spheres are separated far apart.

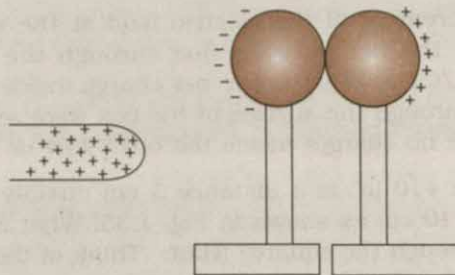


Fig. 1.36

- 1.30** An oil drop of 12 excess electrons is held stationary under a constant electric field of $2.55 \times 10^4 \text{ NC}^{-1}$ in Millikan's oil drop experiment. The density of the oil is 1.26 g cm^{-3} . Estimate the radius of the drop. ($g = 9.81 \text{ m s}^{-2}$; $e = 1.60 \times 10^{-19} \text{ C}$).
- 1.31** Which among the curves shown in Fig. 1.37 cannot possibly represent electrostatic field lines?

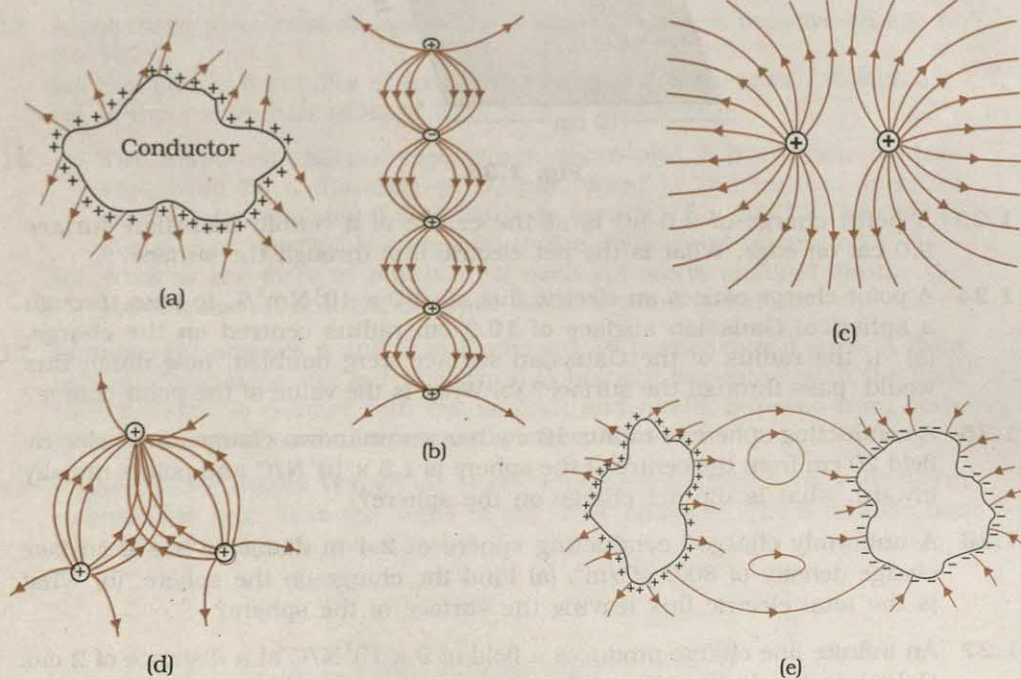


Fig. 1.37

- 1.32** In a certain region of space, electric field is along the z -direction throughout. The magnitude of electric field is, however, not constant but increases

uniformly along the positive z -direction, at the rate of 10^5 NC^{-1} per metre. What are the force and torque experienced by a system having a total dipole moment equal to 10^{-7} Cm in the negative z -direction ?

- 1.33** (a) A conductor A with a cavity as shown in Fig. 1.38(a) is given a charge Q . Show that the entire charge must appear on the outer surface of the conductor. (b) Another conductor B with charge q is inserted into the cavity keeping B insulated from A. Show that the total charge on the outside surface of A is $Q + q$ [Fig. 1.38(b)]. (c) A sensitive instrument is to be shielded from the strong electrostatic fields in its environment. Suggest a possible way.

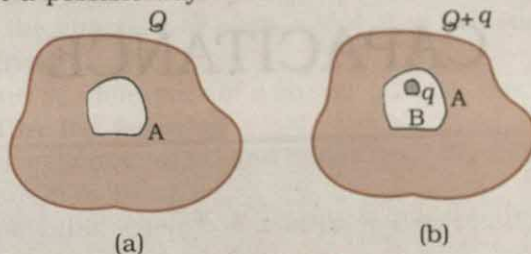
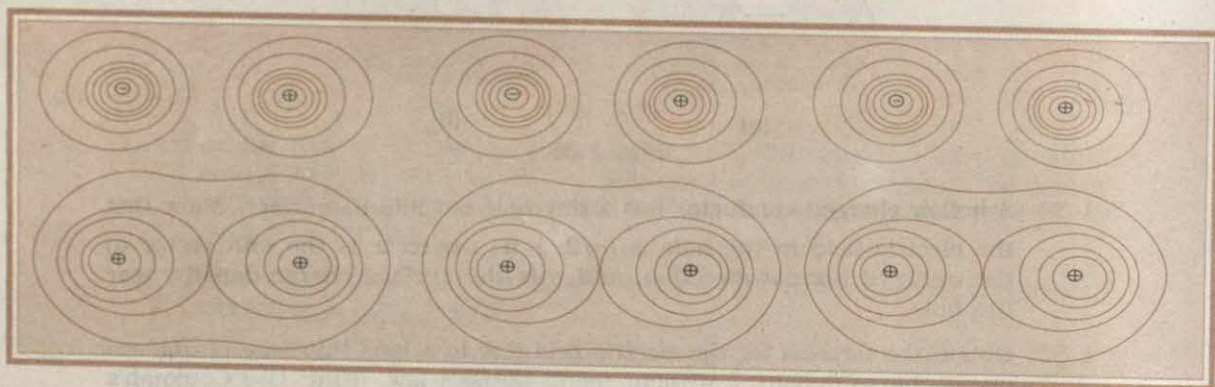


Fig. 1.38

- 1.34** A hollow charged conductor has a tiny hole cut into its surface. Show that the electric field in the hole is $(\sigma/2\epsilon_0) \hat{n}$, where \hat{n} is the unit vector in the outward normal direction, and σ is the surface charge density near the hole.
- 1.35** Obtain the formula for the electric field due to a long thin wire of uniform linear charge density λ without using Gauss's law. [Hint: Use Coulomb's law directly and evaluate the necessary integral.]
- 1.36** It is now believed that protons and neutrons (which constitute nuclei of ordinary matter) are themselves built out of more elementary units called quarks. A proton and a neutron consist of three quarks each. Two types of quarks, the so called 'up' quark (denoted by u) of charge $+(2/3)e$, and the 'down' quark (denoted by d) of charge $(-1/3)e$, together with electrons build up ordinary matter. (Quarks of other types have also been found which give rise to different unusual varieties of matter.) Suggest a possible quark composition of a proton and neutron.
- 1.37** (a) Consider an arbitrary electrostatic field configuration. A small test charge is placed at a null point (i.e., where $\mathbf{E} = 0$) of the configuration. Show that the equilibrium of the test charge is necessarily unstable.
 (b) Verify this result for the simple configuration of two charges of the same magnitude and sign placed a certain distance apart.

CHAPTER TWO

ELECTROSTATIC POTENTIAL AND CAPACITANCE



2.1 INTRODUCTION

In Chapters 6 and 8 (Class XI), the notion of potential energy was introduced. When an external force does work in taking a body from one point to another against a force like spring force or gravitational force, that work gets stored up as potential energy of the body. When the external force is removed, the body moves, gaining kinetic energy and losing an equal amount of potential energy. The sum of kinetic and potential energies is thus conserved. Forces of this kind are called conservative forces. Spring force and gravitational force are examples of a conservative force.

Coulomb force between two charges is also a conservative force. This is not surprising since it is mathematically similar to the gravitational force; both have inverse-square dependence on distance and differ mainly in the proportionality constants – the masses in the gravitational law are replaced by charges in Coulomb's law. Thus, like the potential energy of a mass in a gravitational field, we can define electrostatic potential energy of a charge in an electrostatic field.

Consider an electrostatic field \mathbf{E} due to some charge configuration. First, for simplicity, consider the field \mathbf{E} due to a charge Q placed at the origin. Now, imagine that we bring a test charge q from a point R to a point P against the repulsive force on it due to the charge Q . With reference to Fig. 2.1, this will happen if Q and q are both positive or both negative. For definiteness, let us take $Q, q > 0$.

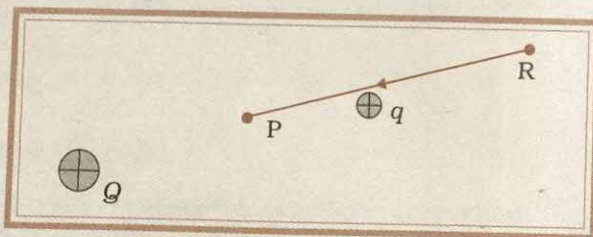


Fig. 2.1 A test charge $q (> 0)$ is moved from the point R to the point P against the repulsive force on it by the charge $Q (> 0)$ placed at the origin.

Two remarks may be made here. First, we assume that the test charge q is so small that it does not disturb the original configuration, namely the charge Q at the origin (or else, we keep Q fixed at the origin by some unspecified force). Second, in bringing the charge q from R to P, we apply an external force just enough to counter the repulsive electric force. This means there is no finite force or acceleration of the charge q when it is brought from R to P – it is brought with infinitesimally slow speed. In this situation, work done by the external force is the negative of the work done by the electric force, and gets fully stored in the form of potential energy of the charge q . If the external force is removed on reaching P, the electric force will take the charge back to R – the stored energy (potential energy) at P is used to provide kinetic energy to the charge q in such a way that the sum of kinetic and potential energies is conserved.

The potential energy difference of a charge q in the electric field due to the charge Q is then defined by the following:

Work done by external force (equal and opposite to the electrostatic force) in bringing charge q from point R to point P

= Difference in potential energy of charge q between final and initial points. In symbols

$$W'_{RP} = V_P - V_R \quad (2.1)$$

where V_P and V_R are the potential energies of the charge q at P and R, respectively and W'_{RP} represents the *work done by the external force* from R to P. Now, since the external force, as noted already, is always to be taken equal and opposite to the electrical force, Eq.(2.1) is equivalent to

$$W_{RP} = V_R - V_P \quad (2.2)$$

where the symbol W_{RP} [without the prime (') on W] stands for the *work done by the electrical force on q due to Q* . It is also referred to as the work done on the charge q by the field due to Q . In Fig. 2.1, $W'_{RP} > 0$, so $W_{RP} < 0$. But the defining Eq. (2.2) is true whatever the signs of Q , q and the work done (externally or by the field).

Equation (2.2), in fact, provides a general definition of potential energy difference of any charge q in a field produced by an arbitrary charge configuration. That is, if instead of a single charge Q , we have a system of charges q_1, q_2, \dots, q_n , Eq. (2.2) still provides the definition of potential energy difference of a charge q in the field produced by the charges q_1, q_2, \dots, q_n .

Two important comments may be made at this stage:

1. The right side of Eq. (2.2) depends only on the initial and final positions of the charge. It means that the work done by an electrostatic field in moving a charge from one point to another depends only on the initial and final points and is independent of the path taken to go from one point to the other (Fig. 2.2). This is a fundamental characteristic of a conservative force. The concept of potential energy would not be meaningful if work depended on the path. The path-independence of work done by an electrostatic field can be proved using Coulomb's law. We omit this proof here.
2. Equation (2.2) defines '*potential energy difference*' in terms of the physically meaningful quantity 'work'. Clearly, potential energy so defined is undetermined to within an additive constant. What this means is that the absolute value of potential energy is not physically significant; it is only the difference of potential

energy that is significant. We can always add an arbitrary constant α to potential energy at every point, since this will not change the potential energy difference:

$$(V_P + \alpha) - (V_R + \alpha) = V_P - V_R$$

Put it differently, the point where potential energy is zero is a matter of choice. A convenient choice is to have electrostatic potential energy zero at infinity. With this choice, if we take the point R at infinity, we get from Eq. (2.1):

$$W'_{\infty P} = V_P - V_{\infty} = V_P \quad (2.3)$$

Since the point P is arbitrary, Eq. (2.3) provides us with a definition of potential energy of charge q at any point: Potential energy of charge q at a point (in the presence of field due to any charge configuration) is the work done by the external force (equal and opposite to the electric force) in bringing the charge q from infinity to that point.



Conte Alessandro Volta (1745-1827)

- Italian physicist, professor at Pavia. Volta established that the “animal electricity” observed by Luigi Galvani, 1737-1798, in experiments with frog muscle tissue placed in contact with dissimilar metals, was not due to any exceptional property of animal tissues but was also generated whenever any wet body was sandwiched between dissimilar metals. This led him to develop the first *voltaic pile*, or battery, consisting of a large stack of moist disks of cardboard (electrolyte) sandwiched between disks of metal (electrodes).

2.2 ELECTROSTATIC POTENTIAL

Consider any general charge configuration. We defined potential energy of a test charge q in terms of the work done on the charge q . This work is obviously proportional to q , since the force at any point is $q\mathbf{E}$, where \mathbf{E} is the electric field at that point due to the given charge configuration. It is, therefore, convenient to divide the work by the amount of charge q , so that the resulting quantity is independent of q . In other words, work done per unit test charge is characteristic of the electric field associated with the charge configuration. This leads to the idea of electrostatic potential ϕ due to a given charge configuration. From Eq. (2.1), we get:

Work done by external force in bringing a unit positive charge from point R to point P

$$= \phi_P - \phi_R \quad (2.4)$$

where ϕ_P and ϕ_R are the electrostatic potential at P and R, respectively. Note, as before, that it is not the absolute value of potential but the potential difference that is physically significant. If, as before, we choose the potential to be zero at infinity, Eq. (2.4) implies:

Work done by external force in bringing a unit positive charge from infinity to a point
= electrostatic potential (ϕ) at that point. (2.5)

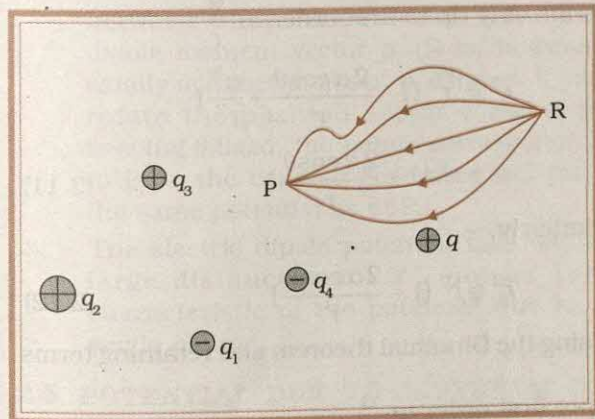


Fig. 2.2 Work done on a test charge q by the electrostatic field due to any charge configuration is independent of the path, and depends only on its initial and final positions.

The qualifying remarks made earlier for potential energy also apply to the definition of potential. To obtain work per unit test charge, we should take an infinitesimal test charge δq ,

obtain the work $\delta W'$ in bringing it from infinity to the point and determine the ratio $\delta W'/\delta q$. Also, the external force at every point of the path is to be equal and opposite to the electrostatic force on the test charge at that point.

2.3 POTENTIAL DUE TO A POINT CHARGE

Consider a point charge Q at the origin (Fig. 2.3). For definiteness, take Q to be positive. We wish to determine the potential at any point P with position vector \mathbf{r} from the origin. For that we must calculate the work done in bringing a unit positive test charge from infinity to the point P. For $Q > 0$, work done against the repulsive force on the test charge is positive. Since work done is independent of path, we choose a convenient path: along the radial direction from infinity to the point P.

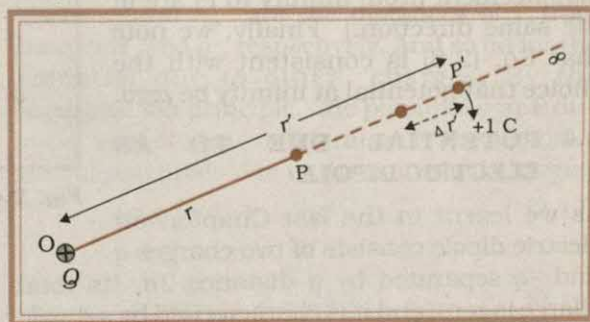


Fig. 2.3 Work done in bringing a unit positive test charge from infinity to the point P, against the repulsive force of charge Q ($Q > 0$), is the potential at P due to the charge Q .

At some intermediate point P' on the path, the electrostatic force on a unit positive charge is

$$\frac{Q \times 1}{4\pi\epsilon_0 r'^2} \hat{\mathbf{r}}'$$

where $\hat{\mathbf{r}}'$ is the unit vector along OP' . Work done against this force from r' to $r' + \Delta r'$:

$$\Delta W' = -\frac{Q}{4\pi\epsilon_0 r'^2} \Delta r' \quad (2.6)$$

The negative sign appears because for $\Delta r' < 0$, $\Delta W'$ is positive. Total work done (W') by the external force is obtained by integrating Eq. (2.6) from $r' = \infty$ to $r' = r$:

$$W' = -\int_{\infty}^r \frac{Q}{4\pi\epsilon_0 r'^2} dr' = \left. \frac{Q}{4\pi\epsilon_0 r'} \right|_{\infty}^r = \frac{Q}{4\pi\epsilon_0 r} \quad (2.7)$$

This, by definition, is the potential at P due to the charge Q :

$$\phi = \frac{Q}{4\pi\epsilon_0 r} \quad (2.8)$$

Equation (2.8) is true for any sign of the charge Q , though we considered $Q > 0$ in its derivation. For $Q < 0$, $\phi < 0$, i.e., work done (by the external force) per unit positive test charge from infinity to the point is negative. This is equivalent to saying that work done by the electrostatic force in bringing the unit positive charge from infinity to the point P is positive. [This is as it should be, since for $Q < 0$, the force on a unit positive test charge is attractive, so that the electrostatic force and the displacement (from infinity to P) are in the same direction.] Finally, we note that Eq. (2.8) is consistent with the choice that potential at infinity be zero.

2.4 POTENTIAL DUE TO AN ELECTRIC DIPOLE

As we learnt in the last Chapter, an electric dipole consists of two charges q and $-q$ separated by a distance $2a$. Its total charge is zero, and it is characterised by a dipole moment vector \mathbf{p} whose magnitude is $q \times 2a$ and which points in the direction from $-q$ to q (Fig. 2.4). We also saw that the electric field of a dipole at a point with position vector \mathbf{r} depends not just on the magnitude r , but also on the angle between \mathbf{r} and \mathbf{p} . Further, the field falls off, at large distance, not as $1/r^2$ (typical of field due to a single charge) but as $1/r^3$. Here we determine the electric potential due to a dipole and contrast it with the potential due to a single charge.

As before, we take the origin at the centre of the dipole. Now we know that the electric field obeys the superposition principle. Since potential is related to the work done by the field, electrostatic potential also follows the superposition principle. Thus, potential due to the dipole is the sum of potentials due to the charges q and $-q$:

$$\phi = \frac{1}{4\pi\epsilon_0} \left(\frac{q}{r_1} - \frac{q}{r_2} \right) \quad (2.9)$$

where r_1 and r_2 are the distances of the point P from q and $-q$, respectively.

Now, by geometry,

$$r_1^2 = r^2 + a^2 - 2ar \cos \theta$$

$$r_2^2 = r^2 + a^2 + 2ar \cos \theta \quad (2.10)$$

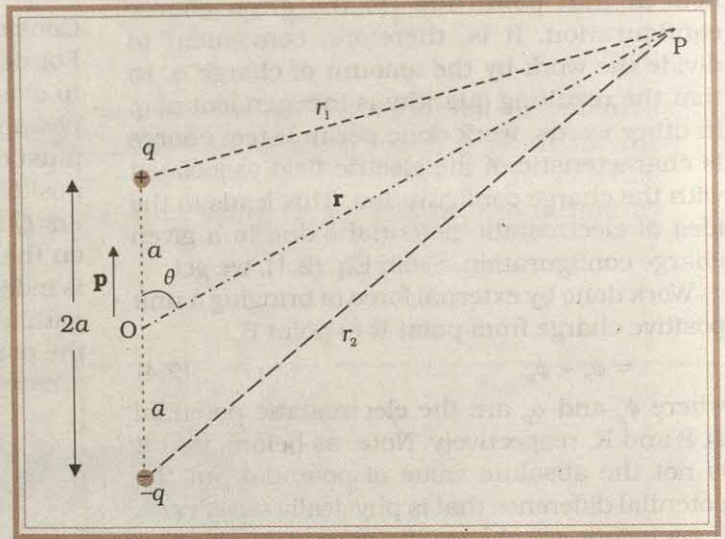


Fig. 2.4 Quantities involved in the calculation of potential due to a dipole.

We take r much greater than a ($r \gg a$) and retain terms only up to first order in $\frac{a}{r}$:

$$\begin{aligned} r_1^2 &= r^2 \left(1 - \frac{2a \cos \theta}{r} + \frac{a^2}{r^2} \right) \\ &\approx r^2 \left(1 - \frac{2a \cos \theta}{r} \right) \end{aligned} \quad (2.11)$$

Similarly,

$$r_2^2 \approx r^2 \left(1 + \frac{2a \cos \theta}{r} \right) \quad (2.12)$$

Using the Binomial theorem and retaining terms

upto first order in $\frac{a}{r}$,

$$\frac{1}{r_1} \approx \frac{1}{r} \left(1 - \frac{2a \cos \theta}{r} \right)^{-\frac{1}{2}} \approx \frac{1}{r} \left(1 + \frac{a}{r} \cos \theta \right) \quad (2.13a)$$

$$\frac{1}{r_2} \approx \frac{1}{r} \left(1 + \frac{2a \cos \theta}{r} \right)^{-\frac{1}{2}} \approx \frac{1}{r} \left(1 - \frac{a}{r} \cos \theta \right) \quad (2.13b)$$

Using Eqs. (2.9) and (2.13) and $p = 2qa$, we get

$$\phi = \frac{q}{4\pi\epsilon_0} \frac{2a \cos\theta}{r^2} = \frac{p \cos\theta}{4\pi\epsilon_0 r^2} \quad (2.14)$$

Now $p \cos\theta = \mathbf{p} \cdot \hat{\mathbf{r}}$

where $\hat{\mathbf{r}}$ is the unit vector along the position vector \mathbf{OP} .

The electric potential of a dipole is then given by

$$\phi = \frac{1}{4\pi\epsilon_0} \frac{\mathbf{p} \cdot \hat{\mathbf{r}}}{r^2}; \quad (r \gg a) \quad (2.15)$$

Equation (2.15) is, as indicated, approximately true only for distances large compared to the size of the dipole, so that higher order terms in a/r are negligible. For a point dipole \mathbf{p} at the origin, Eq. (2.15) is, however, exact.

From Eq. (2.15), potential on the dipole axis ($\theta = 0, \pi$) is given by

$$\phi = \pm \frac{1}{4\pi\epsilon_0} \frac{p}{r^2} \quad (2.16)$$

(+ sign for $\theta = 0$, -sign for $\theta = \pi$). Potential in the equatorial plane ($\theta = \pi/2$) is zero.

The important contrasting features of electric potential of a dipole from that due to a single charge are clear from Eqs. (2.8) and (2.15):

1. The potential due to a dipole depends not just on distance r but also on the angle between the position vector \mathbf{r} and the dipole moment vector \mathbf{p} . (It is, however, axially symmetric about \mathbf{p} . That is, if you rotate the position vector \mathbf{r} about \mathbf{p} , keeping θ fixed, the points corresponding to P on the cone so generated will have the same potential as at P.)
2. The electric dipole potential falls off, at large distance, as $1/r^2$, not as $1/r$, characteristic of the potential due to a single charge.

2.5 POTENTIAL DUE TO A SYSTEM OF CHARGES

Consider a system of charges q_1, q_2, \dots, q_n , with position vectors $\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_n$ relative to some origin (Fig. 2.5). The potential ϕ_1 at P due to the charge q_1 is:

$$\phi_1 = \frac{1}{4\pi\epsilon_0} \frac{q_1}{r_{1P}}$$

where r_{1P} is the distance between q_1 and P.

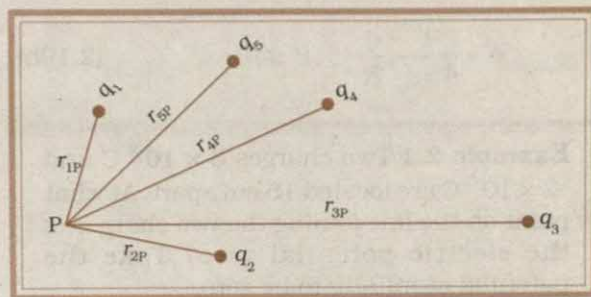


Fig. 2.5 Potential at a point due to a system of charges is the sum of potentials due to individual charges.

Similarly, the potential ϕ_2 at P due to q_2 and ϕ_3 due to q_3 are given by

$$\phi_2 = \frac{1}{4\pi\epsilon_0} \frac{q_2}{r_{2P}}, \quad \phi_3 = \frac{1}{4\pi\epsilon_0} \frac{q_3}{r_{3P}}$$

where r_{2P} and r_{3P} are the distances of P from charges q_2 and q_3 , respectively. And so on for the potential due to other charges. By the superposition principle, the potential ϕ at P due to the total charge configuration is the algebraic sum of potentials due to the individual charges:

$$\phi = \phi_1 + \phi_2 + \dots + \phi_n \quad (2.17)$$

$$= \frac{1}{4\pi\epsilon_0} \left(\frac{q_1}{r_{1P}} + \frac{q_2}{r_{2P}} + \dots + \frac{q_n}{r_{nP}} \right) \quad (2.18)$$

If we have a continuous charge distribution characterised by a charge density ρ , we divide it, as before, into small volume elements each of size ΔV and carrying charge $\rho \Delta V$. We then determine the potential due to each volume element and sum (strictly speaking, integrate) over all such contributions, and thus determine the potential due to the entire distribution.

We have seen in Chapter 1 that for a *uniformly charged spherical shell*, the electric field outside the shell is as if the entire charge is concentrated at the centre. Thus, the potential outside the shell is given by

$$\phi = \frac{1}{4\pi\epsilon_0} \frac{q}{r} \quad (r \geq R) \quad (2.19a)$$

where q is the total charge on the shell and R its radius. The electric field inside the shell is zero. This implies (Section 2.6) that potential is constant inside the shell and, therefore, equals its value at the surface.

$$\phi = \frac{1}{4\pi\epsilon_0} \frac{q}{R} \quad (r \leq R) \quad (2.19b)$$

Example 2.1 Two charges $3 \times 10^{-8} \text{ C}$ and $-2 \times 10^{-8} \text{ C}$ are located 15 cm apart. At what point on the line joining the two charges is the electric potential zero? Take the potential at infinity to be zero.

Answer Let us take the origin O at the location of the positive charge. The line joining the two charges is taken to be the x -axis; the negative charge is taken to be on the right side of the origin (Fig. 2.6).

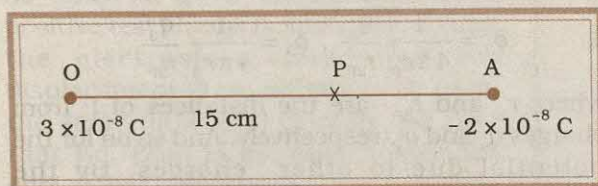


Fig. 2.6 Zero of electric potential for two charges.

Let P be the required point on the x -axis where the potential is zero. If x is the x -coordinate of P, obviously x must be positive. (There is no possibility of potentials due to the two charges adding up to zero for $x < 0$). If x lies between O and A, we have

$$\frac{1}{4\pi\epsilon_0} \left[\frac{3 \times 10^{-8}}{x \times 10^{-2}} - \frac{2 \times 10^{-8}}{(15 - x) \times 10^{-2}} \right] = 0$$

where x is in cm. That is,

$$\frac{3}{x} - \frac{2}{15 - x} = 0$$

which gives $x = 9 \text{ cm}$.

If x lies on the extended line OA, the required condition is

$$\frac{3}{x} - \frac{2}{x - 15} = 0$$

which gives

$$x = 45 \text{ cm}$$

Thus, electric potential is zero at 9 cm and 45 cm away from the positive charge on the side of the negative charge. Note that the formula for potential used in the calculation required choosing potential to be zero at infinity.

2.6 EQUIPOTENTIAL SURFACES

An equipotential surface is a surface with a constant value of potential at all points on the surface. For a single charge q , the potential is given by Eq. (2.8):

$$\phi = \frac{1}{4\pi\epsilon_0} \frac{q}{r}$$

This shows that ϕ is constant if r is constant. Thus, equipotential surfaces of a single point charge are concentric spherical surfaces centred at the charge.

Now the electric field lines for a single charge q are radial lines starting from or ending at the charge, depending on whether q is positive or negative. Clearly, the electric field at every point is normal to the equipotential surface passing through that point. This is true in general: for any charge configuration, equipotential surface through a point is normal to the electric field at that point. The proof of this statement is simple.

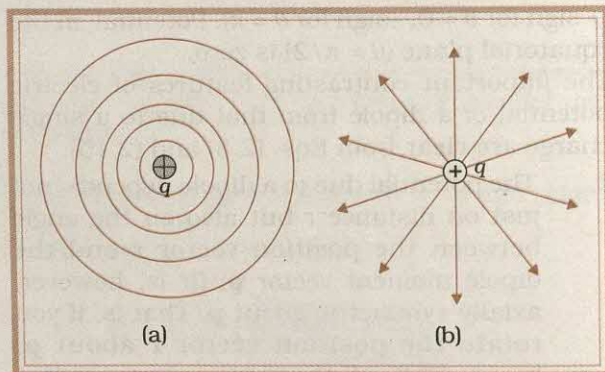


Fig. 2.7 For a single charge q (a) equipotential surfaces are spherical surfaces centred at the charge, and (b) electric field lines are radial, starting from the charge if $q > 0$.

If the field were not normal to the equipotential surface, it would have non-zero component along the surface. To move a unit test charge against the direction of the component of the field, work would have to be done. But this is against the definition of an equipotential surface: there is no potential difference between any two points on the surface and no work is required to move a test charge on the surface. The electric field must, therefore, be normal to the equipotential surface at every point. Equipotential surfaces offer an alternative visual picture from the picture of electric field lines around a charge configuration.

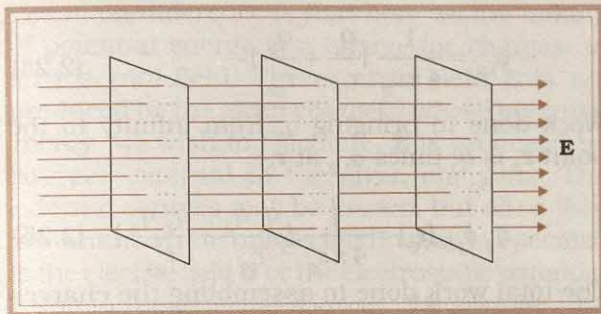


Fig. 2.8 Equipotential surfaces for a uniform electric field.

For a uniform electric field, say, along the x -axis, the equipotential surfaces are planes normal to the x -axis i.e. planes parallel to the y - z plane (Fig. 2.8). Equipotential surfaces for a dipole and its electric field lines are shown in Fig. 2.9.

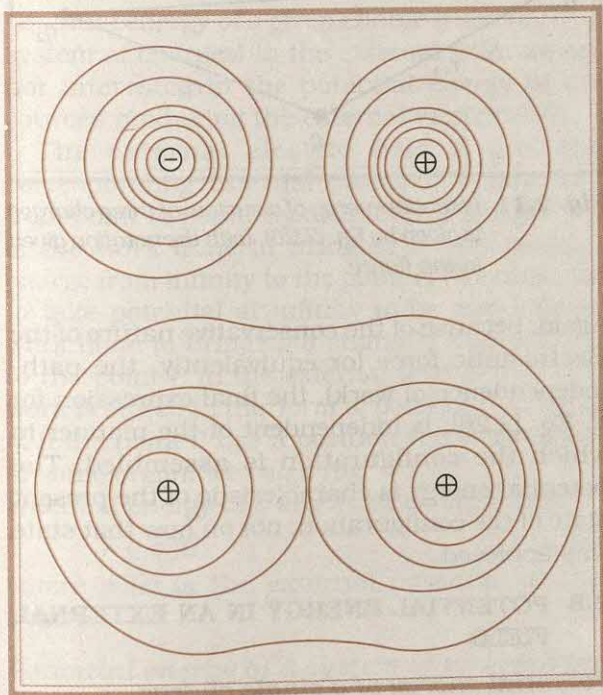


Fig. 2.9 Some equipotential surfaces for the field of a dipole.

Relation between field and potential

Consider two closely spaced equipotential surfaces A and B with potential values ϕ and $\phi + \delta\phi$, where $\delta\phi$ is the change in ϕ in the direction of the electric field \mathbf{E} . Imagine that a unit positive test charge is moved from B to A over the perpendicular distance δl between the two surfaces at point P. Work done on a unit

positive charge (against the electric field) from B to A is $|\mathbf{E}| \delta l$. This work equals the potential difference $\phi_A - \phi_B$. Thus,

$$|\mathbf{E}| \delta l = \phi - (\phi + \delta\phi) = -\delta\phi$$

$$\text{i.e., } |\mathbf{E}| = -\frac{\delta\phi}{\delta l} \quad (2.20)$$

Clearly, $\delta\phi$ is negative, since $|\mathbf{E}|$ is positive. Thus, the direction of electric field is in the direction of decreasing potential. Further, note that the field is in the direction where this decrease is the steepest. We can rewrite Eq. (2.20) as

$$|\mathbf{E}| = -\frac{|\delta\phi|}{\delta l} \quad (2.21)$$

We thus arrive at two important conclusions concerning the relation of electric field and potential. First, electric field is in the direction where the potential decreases steepest. Second, its magnitude is given by the change in the magnitude of potential per unit displacement normal to the equipotential surface at the point.

2.7 POTENTIAL ENERGY OF A SYSTEM OF CHARGES

Consider first the simple case of two charges q_1 and q_2 with position vectors \mathbf{r}_1 and \mathbf{r}_2 relative to some origin. Let us calculate the work done (externally) in building this configuration. This means we consider the charges q_1 and q_2 initially at infinity and determine the work done by an external agency to bring the charges to the given locations. Suppose, first the charge q_1 is brought from infinity to the point \mathbf{r}_1 . There is no external field against which work needs to be done, so work done in bringing q_1 from infinity to \mathbf{r}_1 is zero. This charge produces a potential in space given by

$$\phi_1 = \frac{1}{4\pi\epsilon_0} \frac{q_1}{r_{1P}}$$

where r_{1P} is the distance of any point P in space from the location of q_1 . By the definition of potential, work done in bringing charge q_2 from infinity to the point \mathbf{r}_2 is q_2 times the potential at \mathbf{r}_2 due to q_1 :

$$\text{work done on } q_2 = \frac{1}{4\pi\epsilon_0} \frac{q_1 q_2}{r_{12}}$$

where r_{12} is the distance between points 1 and 2.

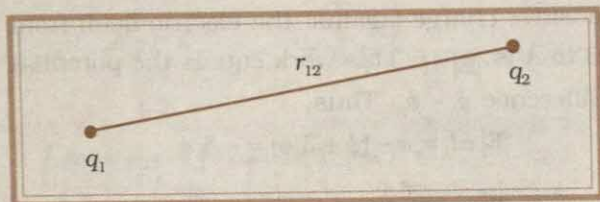


Fig. 2.10 Potential energy of a system of charges q_1 and q_2 is directly proportional to the product of charges and inversely to the distance between them.

Since electrostatic force is conservative, this work gets stored in the form of potential energy of the system. Thus, the potential energy of a system of two charges q_1 and q_2 is :

$$V = \frac{1}{4\pi\epsilon_0} \frac{q_1 q_2}{r_{12}} \quad (2.22)$$

Obviously, if q_2 was brought first to its present location and q_1 brought later, the potential energy V would be the same. More generally, the potential energy expression, Eq. (2.22), is unaltered whatever way the charges are brought to the specified locations, because of path-independence of work for electrostatic force.

Equation (2.22) is true for any sign of q_1 and q_2 . If $q_1 q_2 > 0$, potential energy is positive. This is as expected, since for like charges ($q_1 q_2 > 0$), electrostatic force is repulsive and a positive amount of work is needed to be done against this force to bring the charges from infinity to a finite distance apart. For unlike charges ($q_1 q_2 < 0$), the electrostatic force is attractive. In that case, a positive amount of work is needed against this force to take the charges from the given locations to infinity. In other words, a negative amount of work is needed for the reverse path (from infinity to the present locations), so the potential energy is negative.

Equation (2.22) is easily generalised for a system of any number of charges. Let us calculate the potential energy of a system of three charges q_1, q_2, q_3 located at $\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3$ respectively. To bring q_1 first from infinity to \mathbf{r}_1 , no work is done. Next we bring q_2 from infinity to \mathbf{r}_2 . As before, work done in this step is

$$q_2 \phi_1(\mathbf{r}_2) = \frac{1}{4\pi\epsilon_0} \frac{q_1 q_2}{r_{12}} \quad (2.23)$$

The charges q_1 and q_2 produce potential all around, which at any point P is given by

$$\phi_{1,2} = \frac{1}{4\pi\epsilon_0} \left(\frac{q_1}{r_{1P}} + \frac{q_2}{r_{2P}} \right) \quad (2.24)$$

Work done in bringing q_3 from infinity to the point \mathbf{r}_3 is q_3 times $\phi_{1,2}$ at \mathbf{r}_3 :

$$q_3 \phi_{1,2}(\mathbf{r}_3) = \frac{1}{4\pi\epsilon_0} \left(\frac{q_1 q_3}{r_{13}} + \frac{q_2 q_3}{r_{23}} \right) \quad (2.25)$$

The total work done in assembling the charges at the given locations is obtained by adding the work done in different steps [Eq.(2.23) and Eq. (2.25)].

$$V = \frac{1}{4\pi\epsilon_0} \left[\frac{q_1 q_2}{r_{12}} + \frac{q_1 q_3}{r_{13}} + \frac{q_2 q_3}{r_{23}} \right] \quad (2.26)$$

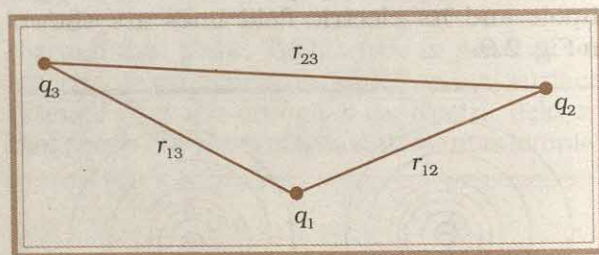


Fig. 2.11 Potential energy of a system of three charges is given by Eq. (2.26), with the notation given in the figure.

Again, because of the conservative nature of the electrostatic force (or equivalently, the path-independence of work), the final expression for V , Eq. (2.26), is independent of the manner in which the configuration is assembled. The potential energy is characteristic of the present state of the configuration, not on how that state was achieved.

2.8 POTENTIAL ENERGY IN AN EXTERNAL FIELD

Potential energy of a single charge

In Section 2.7, the source of the electric field was specified – the charges and their locations – and the potential energy of the system of those charges was determined. In this section, we ask a related but distinct question. What is the potential energy of a charge q in a given field? This question was, in fact, the starting point that led us to the notion of electrostatic potential (Sections 2.1 and 2.2). But we address that question here again to clarify in what way it is different from the discussion in section 2.7.

The main difference is that here we are talking of potential energy of a charge (or charges) in an *external* field. The external field \mathbf{E} is *not* produced by the given charge(s) whose potential energy we wish to calculate. \mathbf{E} is produced by sources external to the given charge(s). The external sources may be known, but often they are unknown or unspecified; what is specified is the electric field \mathbf{E} or the electrostatic potential ϕ due to the external sources. We assume that the charge q does not significantly affect the sources producing the external field. This is true if q is very small, or the external sources are held fixed by other unspecified forces. Even if q is finite, its influence on the external sources may still be ignored in the situation when very strong sources far away at infinity produce a finite field \mathbf{E} in the region of interest. Note again that we are interested in determining the potential energy of a given charge q (and later, a system of charges) in the external field; we are not interested in the potential energy of the sources producing the external electric field.

The external electric field \mathbf{E} and the corresponding external potential ϕ may vary from point to point. By definition, ϕ at a point P is the work done in bringing a unit positive charge from infinity to the point P. (We continue to take potential at infinity to be zero.) Thus, work done in bringing a charge q from infinity to the point P in the external field is $q\phi$. This work is stored in the form of potential energy of q . If the point P has a position vector \mathbf{r} relative to some origin, we can write:

$$\begin{aligned} \text{Potential energy of } q \text{ at } \mathbf{r} \text{ in an external field} \\ = q\phi(\mathbf{r}) \end{aligned} \quad (2.27)$$

where $\phi(\mathbf{r})$ is the external potential at the point \mathbf{r} .

Potential energy of a system of two charges in an external field

Next, we ask: what is the potential energy of a system of two charges q_1 and q_2 located at \mathbf{r}_1 and \mathbf{r}_2 , respectively, in an external field? To calculate the work done in building this configuration, let us imagine bringing the charge q_1 first from infinity to \mathbf{r}_1 . Work done in this step is $q_1\phi(\mathbf{r}_1)$, using Eq. (2.27). Next, we consider the work done in bringing q_2 to \mathbf{r}_2 . In this step, work is done not only against the external field \mathbf{E} but also against the field due to q_1 :

$$\begin{aligned} \text{Work done on } q_2 \text{ against the external field} \\ = q_2\phi(\mathbf{r}_2) \end{aligned}$$

$$\begin{aligned} \text{Work done on } q_2 \text{ against the field due to } q_1 \\ = \frac{q_1 q_2}{4\pi\epsilon_0 r_{12}} \end{aligned}$$

where r_{12} is the distance between q_1 and q_2 . We have made use of Eqs. (2.27) and (2.22) above. By the superposition principle for fields, we add up the work done on q_2 against the two fields (\mathbf{E} and that due to q_1):

Work done in bringing q_2 at \mathbf{r}_2

$$= q_2\phi(\mathbf{r}_2) + \frac{q_1 q_2}{4\pi\epsilon_0 r_{12}}$$

The total work in bringing q_1 at \mathbf{r}_1 and q_2 at \mathbf{r}_2 is independent of the manner in which we build the configuration. (Remember, path-independence of work against electrostatic fields.) The total work appears as potential energy of the system. Thus, potential energy of two charges q_1, q_2 located at $\mathbf{r}_1, \mathbf{r}_2$ in an external field

$$= q_1\phi(\mathbf{r}_1) + q_2\phi(\mathbf{r}_2) + \frac{q_1 q_2}{4\pi\epsilon_0 r_{12}} \quad (2.28)$$

Potential energy of a dipole

The case of a dipole in a uniform external field is an interesting application of Eq. (2.28). For a dipole $q_1 = q, q_2 = -q$.

Suppose the external uniform field \mathbf{E} is in the x -direction. Let the origin be the centre of the dipole. Potential energy of the dipole

$$= q[\phi(\mathbf{r}_1) - \phi(\mathbf{r}_2)] - \frac{q^2}{4\pi\epsilon_0 \times 2a} \quad (2.29)$$

The two point charges q_1 and q_2 at locations \mathbf{r}_1 and \mathbf{r}_2 of the two charges have been numbered 1 and 2 for convenience.

Now potential difference between 1 and 2 equals the work done in bringing a unit positive charge against the field from 2 to 1. Since work done is force multiplied by the displacement parallel to the force,

$$\phi(\mathbf{r}_1) - \phi(\mathbf{r}_2) = -E \times 2a \cos \theta \quad (2.30)$$

The negative sign in Eq. (2.30) agrees with the fact that potential decreases in the direction of the field. Thus potential energy of a dipole in a uniform external field \mathbf{E}

$$= -q \times E \times 2a \cos \theta - \frac{q^2}{4\pi\epsilon_0 \times 2a} \quad (2.31)$$

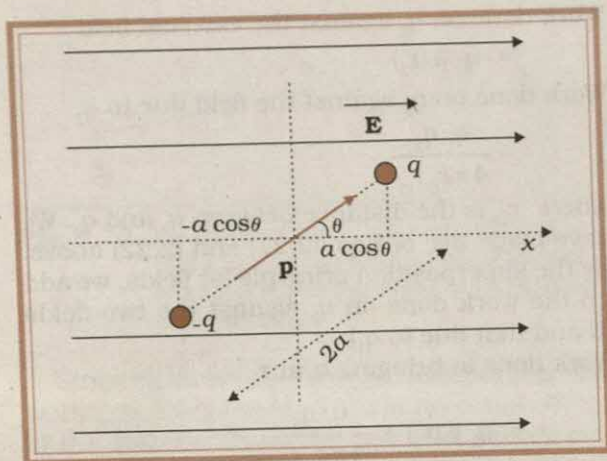


Fig. 2.12 Potential energy of a dipole in a uniform external field.

Now for a given dipole, the second term is only a constant*. Since a constant is insignificant for potential energy, we can drop the second term in Eq. (2.31) and write :

Potential energy of a dipole in a uniform \mathbf{E}

$$= -q \times 2aE \cos \theta = -pE \cos \theta \quad (2.32)$$

$$= -\mathbf{p} \cdot \mathbf{E} \quad (2.33)$$

Clearly, the dipole has minimum potential energy when it is aligned with the field. Now recall the result that a dipole in a uniform field experiences a torque ($\boldsymbol{\tau} = \mathbf{p} \times \mathbf{E}$) which tends to align the dipole with the field direction. Thus, if the dipole can dissipate its potential energy (in the form of heat to the surroundings), the torque will align the dipole with the direction of the external field, bringing its potential energy to the minimum.

Example 2.2 Four charges are arranged at the corners of a square ABCD of side d as shown in Fig. 2.13. (a) Find the work required to put together this arrangement. (b) A charge q_0 is brought to the centre E of the square, the four charges being held fixed at its corners. How much extra work is needed to do this ?

Answer

- (a) Since the work done depends on the final arrangement of the charges, and not on how they are put together, we calculate the work needed for one way of putting the charges at A, B, C and D.

Suppose, first the charge $+q$ is brought to A, and then the charges $-q$, $+q$, and $-q$ are brought to B, C and D respectively. The total work needed can be calculated in steps :

- (i) Work needed to bring charge $+q$ to A when no charge is present elsewhere: this is zero.
- (ii) Work needed to bring charge $-q$ to B when $+q$ is at A. This is given by (charge at B) \times (electrostatic potential at B due to charge $+q$ at A)

$$= -q \times \left(\frac{q}{4\pi\epsilon_0 d} \right) = -\frac{q^2}{4\pi\epsilon_0 d}$$

- (iii) Work needed to bring charge $+q$ to C when $+q$ is at A and $-q$ is at B. This is given by (charge at C) \times (potential at C due to charges at A and B)

$$= +q \left(\frac{+q}{4\pi\epsilon_0 d\sqrt{2}} + \frac{-q}{4\pi\epsilon_0 d} \right)$$

$$= \frac{-q^2}{4\pi\epsilon_0 d} \left(1 - \frac{1}{\sqrt{2}} \right)$$

- (iv) Work needed to bring $-q$ to D when $+q$ is at A, $-q$ at B, and $+q$ at C. This is given by (charge at D) \times (potential at D due to charges at A, B and C)

$$= -q \left(\frac{+q}{4\pi\epsilon_0 d} + \frac{(-q)}{4\pi\epsilon_0 d\sqrt{2}} + \frac{q}{4\pi\epsilon_0 d} \right)$$

$$= -\frac{q^2}{4\pi\epsilon_0 d} \left(2 - \frac{1}{\sqrt{2}} \right)$$

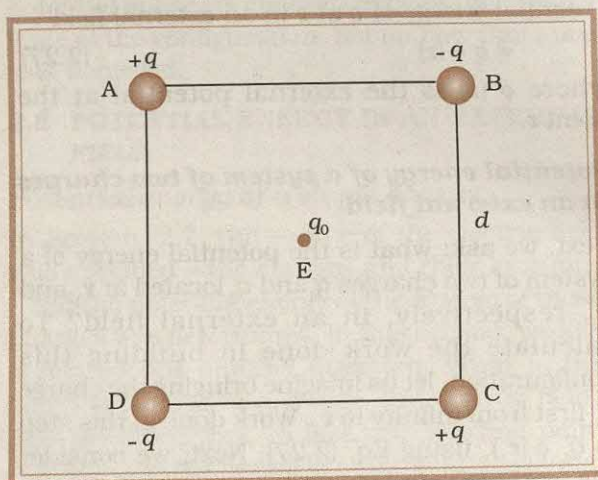



Fig. 2.13 The system of charges for Example 2.2.

*This is not true for a dipole induced by the field. We are considering a permanent dipole.

Add the work done in steps (i), (ii), (iii) and (iv). The total work required is

$$\begin{aligned}
 &= \frac{-q^2}{4\pi\epsilon_0 d} \left\{ (0) + (1) + \left(1 - \frac{1}{\sqrt{2}}\right) + \left(2 - \frac{1}{\sqrt{2}}\right) \right\} \\
 &= -\frac{q^2}{4\pi\epsilon_0 d} (4 - \sqrt{2})
 \end{aligned}$$

The work done depends only on the arrangements of the charges, and not on how they are assembled. By definition, this is the total electrostatic energy of the charges.

- (b) The extra work necessary to bring a charge q_0 to the point E when the four charges are at A, B, C and D is $q_0 \times$ (electrostatic potential at E due to the charges at A, B, C and D). The electrostatic potential at E is clearly zero since potential due to A and C is cancelled by that due to B and D. Hence no work is required to bring any charge to the point E. 

2.9 ELECTROSTATICS OF CONDUCTORS

Conductors and insulators were described briefly in Chapter 1. Conductors contain mobile charge carriers. In metallic conductors, these charge carriers are electrons. In a metal, the outer (valence) electrons part from their atoms and are free to move. The free electrons are a kind of 'gas'; they collide with each other and with the ions, and move randomly in different directions. In an external electric field, they drift against the direction of the field. The positive ions made up of the nuclei and the bound electrons remain held in their fixed positions. In electrolytic conductors, the charge carriers are both positive and negative ions; but the situation in this case is more involved – the movement of the charge carriers is affected both by the external electric field as also by the so-called chemical forces (see Chapter 3). We shall restrict our discussion to metallic solid conductors. Let us note some important results regarding the electrostatics of conductors:

1. *Inside a conductor, electrostatic field is zero.*

Consider a conductor, neutral or charged. There may also be an external electrostatic field. In the static situation, when there is no current inside or on the surface of the conductor, the

electric field is zero everywhere inside the conductor. This fact can be taken as the defining property of a conductor. A conductor has free electrons. As long as electric field is not zero, the free charge carriers would experience force and drift. In the static situation, the free charges have so distributed themselves that the electric field is zero everywhere inside. *Electrostatic field is zero inside a conductor.*

2. *At the surface of a charged conductor, electrostatic field must be normal to the surface at every point.*

If \mathbf{E} is not normal to the surface, it would have some non-zero component along the surface. Free charges on the surface of the conductor would then experience force and move. In the static situation, therefore, \mathbf{E} should have no tangential component. Thus *electrostatic field at the surface of a charged conductor must be normal to the surface at every point.* (For a conductor without any surface charge, field is zero even at the surface.) See result 5.

3. *The interior of a conductor can have no excess charge in the static situation.*

A neutral conductor has equal amounts of positive and negative charges in every small volume or surface element. When the conductor is charged, the excess charge can reside only on the surface in the static situation. This follows from Gauss's theorem. Consider any arbitrary volume element V inside a conductor. On the closed surface S bounding the volume element V , electrostatic field is zero. Thus the total electric flux through S is zero. Hence, by Gauss's theorem, there is no net charge enclosed by S . But the surface S can be made as small as you like, i.e., the volume V can be made vanishingly small. This means there is no net charge at any point inside the conductor, and any excess charge must reside at the surface.

4. *Electrostatic potential is constant throughout the volume of a conductor and has the same value (as inside) on its surface.*

This follows from results 1 and 2 above. Since $\mathbf{E} = 0$ inside the conductor and has no tangential component on the surface, no work is done on moving a small test charge within the conductor and on its surface. That is, there

is no potential difference between any two points inside or on the surface of the conductor. Hence the result. If the conductor is charged, electric field normal to the surface exists; this means potential will be different for the surface and a point just outside the surface.

In a system of conductors of arbitrary size, shape and charge configuration, each conductor is characterised by a constant value of potential, but this constant may differ from one conductor to the other.

5. Electric field at the surface of a charged conductor

$$\mathbf{E} = \frac{\sigma}{\epsilon_0} \hat{\mathbf{n}} \quad (2.34)$$

where σ is the surface charge density and $\hat{\mathbf{n}}$ is a unit vector normal to the surface in the outward direction.

To derive the result, choose a pill box (a short cylinder) as the Gaussian surface about any point P on the surface, as shown in Fig. 2.14. The pill box is partly inside and partly outside the surface of the conductor. It has a small area of cross section δS and negligible height.

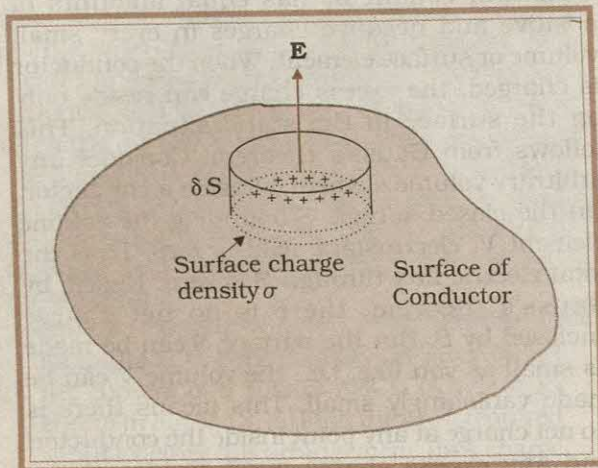


Fig. 2.14 The Gaussian surface (a pill box) chosen to derive Eq. (2.34) for electric field at the surface of a charged conductor.

Just inside the surface, the electrostatic field is zero; just outside, the field is normal to the surface. Thus, the contribution to the total flux through the pill box comes only from the outside (circular) cross-section of the pill box. This equals $\pm E \delta S$ (+ for $\sigma > 0$, - for $\sigma < 0$), since over the small area δS , \mathbf{E} may be considered constant

and \mathbf{E} and $\delta \mathbf{S}$ are parallel or antiparallel. The charge enclosed by the pill box is $\sigma \delta S$. By Gauss' theorem.

$$E \delta S = \frac{|\sigma| \delta S}{\epsilon_0}$$

$$E = \frac{|\sigma|}{\epsilon_0} \quad (2.35)$$

Including the fact that electric field is normal to the surface, we get the vector relation, Eq. (2.34). Note that Eq. (2.34) is true for both signs of σ . For $\sigma > 0$, electric field is normal to the surface outward; for $\sigma < 0$, electric field is normal to the surface inward.

6. Electrostatic shielding

Consider a conductor with a cavity, with no charges inside the cavity. A remarkable result is that whatever be the size and shape of the cavity and whatever be the charge on the conductor and the external fields in which it might be placed, the electric field inside the cavity is zero. We have proved a simple case of this result already: the field inside a charged spherical shell is zero. The proof of the result for the shell makes use of the spherical symmetry of the shell (see Chapter 1). But the vanishing of electric field in the (charge-free) cavity of a conductor is, as mentioned above, a very general result. A related result is that even if the conductor is charged, or charges are induced on a neutral conductor by an external field, all charges reside only on the outer surface of a conductor with cavity.

The proofs of the results noted in Fig. 2.15 are omitted here, but we note their important

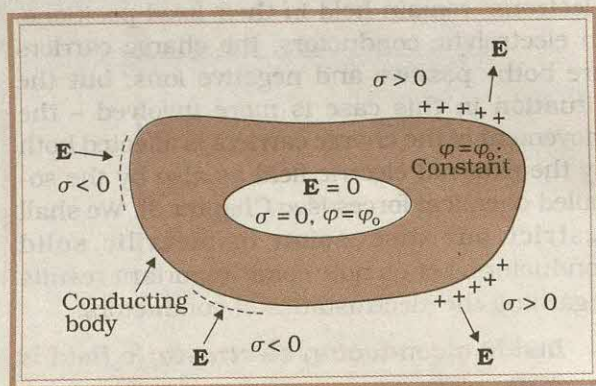


Fig. 2.15 The electric field inside a cavity of any conductor is zero. All charges reside only on the outer surface of a conductor with cavity. (There are no charges placed in the cavity.)

implication. Whatever be the charge and field configuration outside, any cavity in a conductor remains shielded from outside electric influence: the field inside the cavity is always zero. This is known as *electrostatic shielding*. The effect can be made use of in protecting sensitive instruments from outside electrical influences. Fig. 2.16 gives a summary of the important electrostatic properties of a conductor.

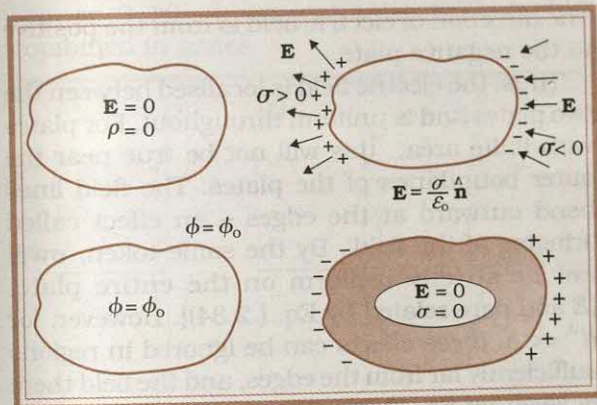


Fig. 2.16 Some important electrostatic properties of a conductor.

2.10 CAPACITORS AND CAPACITANCE

A capacitor is a system of two conductors separated by an insulator (Fig. 2.17). The conductors have charges, say Q_1 and Q_2 , and potentials V_1 and V_2 . Usually, in practice, the two conductors have charges Q and $-Q$, with potential difference $V = V_1 - V_2$ between them. We shall consider only this kind of charge configuration of the capacitor. The conductors may be so charged by connecting them to the two terminals of a battery. Q is called the charge of the capacitor, though this, in fact, is the charge on one of the conductors – the total charge of the capacitor is zero.

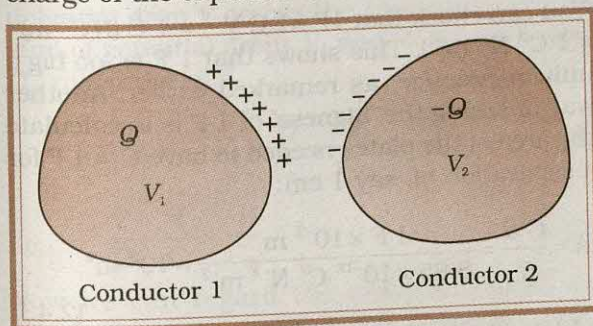


Fig 2.17 A system of two conductors separated by an insulator forms a capacitor.

The electric field in the region between the conductors is proportional to the charge Q . That is, if the charge on the capacitor is, say doubled, the electric field will also be doubled at every point. (This follows from the direct proportionality between field and charge implied by Coulomb's law and the superposition principle.) Now, potential difference V is the work done per unit positive charge in taking a small test charge from the conductor 2 to 1, against the field. Consequently, V is also proportional to Q , and the ratio Q/V is a constant:

$$C = \frac{Q}{V} \quad (2.36)$$

The constant C is called the capacitance of the capacitor. C is independent of Q or V , as said above. The capacitance C depends only on the geometrical configuration (shape, size, separation) of the system of two conductors. [As we shall see later, it also depends on the nature of the insulator (dielectric) separating the two conductors.] The SI unit of capacitance is 1 farad (= 1 coulomb volt⁻¹) or 1 F = 1 C V⁻¹. A capacitor with fixed capacitance is symbolically shown as $\text{||}+$, while the one with variable capacitance is shown as $\text{||}\text{+}$.

Equation (2.36) shows that for large C , V is small for a given Q . This means a capacitor with large capacitance can hold large amount of charge Q at a relatively small V . This is of practical importance. High potential difference implies strong electric field around the conductors. A strong electric field can ionise the surrounding air and accelerate the charges so produced to the oppositely charged plates, thereby neutralising the charge on the capacitor plates, at least partly. In other words, the charge of the capacitor leaks away due to the reduction in insulating power of the intervening medium.

The maximum electric field that a dielectric medium can withstand without break-down (of its insulating property) is called its dielectric strength; for air it is about 3×10^6 Vm⁻¹. For a separation between conductors of the order of 1 cm or so, this field corresponds to a potential difference of 3×10^4 V between the conductors. Thus, for a capacitor to store a large amount of charge without leaking, its capacitance should be high enough so that

the potential difference and hence the electric field do not exceed the break-down limits. Put differently, there is a limit to the amount of charge that can be stored on a given capacitor without significant leaking. In practice, a farad is a very big unit; the more common units are its sub-multiples : $1 \mu\text{F} = 10^{-6} \text{ F}$, $1 \text{ nF} = 10^{-9} \text{ F}$, $1 \text{ pF} = 10^{-12} \text{ F}$, etc. Besides its use in storing charge, a capacitor is a key element of most ac circuits with important functions, as described in Chapter 8.

2.11 THE PARALLEL PLATE CAPACITOR

A parallel plate capacitor consists of two large plane parallel conducting plates separated by a small distance (Fig. 2.18). We first take the intervening medium between the plates to be vacuum. The effect of a dielectric medium between the plates is discussed in Section 2.15. Let A be the area of each plate and d the separation between them. The two plates have charges Q and $-Q$. Since d is much smaller than the linear dimension of the plates ($d^2 \ll A$), we can use the result on electric field by an infinite plane sheet of uniform surface charge density (Section 1.15). Plate 1

has surface charge density $\sigma = \frac{Q}{A}$ and plate 2 has surface charge density $-\sigma$. Using Eq. (1.44), the electric field in different regions is :
Outer region 1:

$$E = \frac{\sigma}{2\epsilon_0} - \frac{\sigma}{2\epsilon_0} = 0 \quad (2.37)$$

Outer region 2 :

$$E = \frac{\sigma}{2\epsilon_0} - \frac{\sigma}{2\epsilon_0} = 0 \quad (2.38)$$

In the inner region between 1 and 2, the electric fields due to the two charged plates add up, giving

$$E = \frac{\sigma}{2\epsilon_0} + \frac{\sigma}{2\epsilon_0} = \frac{\sigma}{\epsilon_0} = \frac{Q}{\epsilon_0 A} \quad (2.39)$$

The direction of electric field is from the positive to the negative plate.

Thus, the electric field is localised between the two plates and is uniform throughout. For plates with finite area, this will not be true near the outer boundaries of the plates. The field lines bend outward at the edges – an effect called ‘fringing of the field’. By the same token, σ will not be strictly uniform on the entire plate. [E and σ are related by Eq. (2.34)]. However, for $d^2 \ll A$, these effects can be ignored in regions sufficiently far from the edges, and the field there is given by Eq. (2.39). Now for uniform electric field, potential difference is simply the electric field times the distance between the plates,

$$\text{i.e., } V = E d = \frac{1}{\epsilon_0} \frac{Q d}{A} \quad (2.40)$$

The capacitance C of the parallel plate capacitor is then

$$C = \frac{Q}{V} = \frac{\epsilon_0 A}{d} \quad (2.41)$$

which, as expected, depends only on the geometry of the system. For typical values like $A = 1 \text{ m}^2$, $d = 1 \text{ mm}$, we get

$$C = \frac{8.85 \times 10^{-12} \text{ C}^2 \text{ N}^{-1} \text{ m}^{-2} \times 1 \text{ m}^2}{10^{-3} \text{ m}} = 8.85 \times 10^{-9} \text{ F} \quad (2.42)$$

(You can check that $1 \text{ F} = 1 \text{ C V}^{-1} = 1 \text{ C (NC}^{-1}\text{m)}^{-1} = 1 \text{ C}^2 \text{ N}^{-1} \text{ m}^{-1}$). This shows that 1 F is too big a unit in practice, as remarked earlier. Another way of seeing the ‘bigness’ of 1 F is to calculate the area of the plates needed to have $C = 1 \text{ F}$ for a separation of, say 1 cm :

$$A = \frac{C d}{\epsilon_0} = \frac{1 \text{ F} \times 10^{-2} \text{ m}}{8.85 \times 10^{-12} \text{ C}^2 \text{ N}^{-1} \text{ m}^{-2}} = 10^9 \text{ m}^2 \quad (2.43)$$

which is a plate about 30 km in length and breadth!

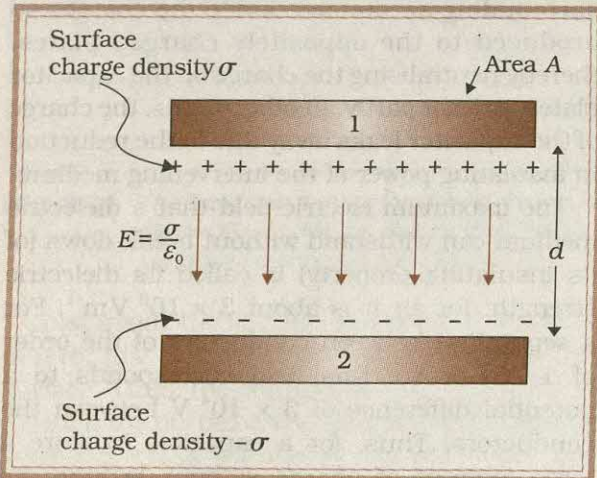


Fig. 2.18 The parallel plate capacitor.

2.12 COMBINATIONS OF CAPACITORS

We can combine several capacitors of capacitance C_1, C_2, \dots, C_n to obtain a system with some effective capacitance C . The effective capacitance depends on the way the individual capacitors are combined. Two simple possibilities are:

Capacitors in series

Figure 2.19 shows capacitors C_1 and C_2 combined in series.

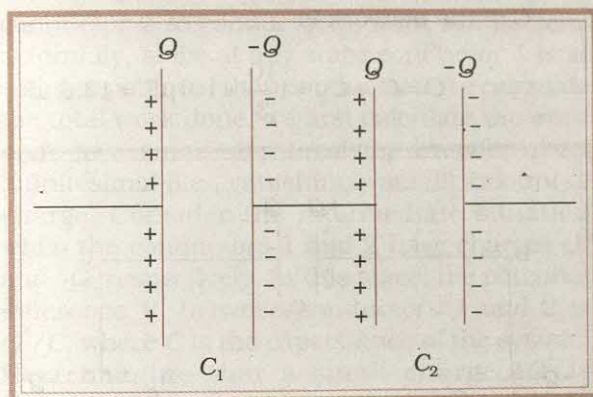


Fig. 2.19 Combination of two capacitors in series.

The left plate of C_1 and the right plate of C_2 are connected to the two terminals of a battery and have charges Q and $-Q$, respectively. It then follows that the right plate of C_1 has charge $-Q$ and the left plate of C_2 has charge Q . If this were not so, the net charge on each capacitor would not be zero. This would result in an electric field in the conductor connecting C_1 and C_2 . Charge would flow until the net charge on both C_1 and C_2 is zero and there is no electric field in the conductor connecting C_1 and C_2 . Thus, in the series combination, charges on the two plates ($\pm Q$) are the same on each capacitor. The total potential drop V across the combination is the sum of potential drops V_1 and V_2 across C_1 and C_2 , respectively:

$$V = V_1 + V_2 = \frac{Q}{C_1} + \frac{Q}{C_2} \quad (2.44)$$

$$\text{i.e., } \frac{V}{Q} = \frac{1}{C_1} + \frac{1}{C_2} \quad (2.45)$$

Now we can regard the combination as an effective capacitor with charge Q and potential difference V . The effective capacitance of the combination is:

$$C = \frac{Q}{V} \quad (2.46)$$

To compare Eq. (2.46) with Eq. (2.45), we invert it:

$$\frac{V}{Q} = \frac{1}{C} \quad (2.47)$$

and thus obtain

$$\frac{1}{C} = \frac{1}{C_1} + \frac{1}{C_2} \quad (2.48)$$

The proof clearly goes through for any number of capacitors arranged in a similar way. Equation (2.44), for n capacitors arranged in series, generalises to

$$V = V_1 + V_2 + \dots + V_n = \frac{Q}{C_1} + \frac{Q}{C_2} + \dots + \frac{Q}{C_n} \quad (2.49)$$

Following the same steps as for the case of two capacitors, we get the general formula for effective capacitance of a series combination of n capacitors:

$$\frac{1}{C} = \frac{1}{C_1} + \frac{1}{C_2} + \dots + \frac{1}{C_n} \quad (2.50)$$

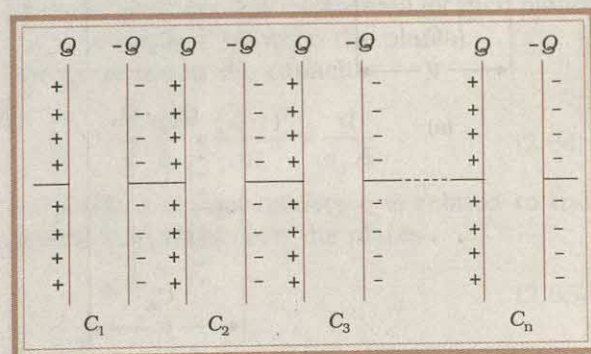


Fig. 2.20 Combination of n capacitors in series.

Capacitors in parallel

Figure 2.21(a) shows two capacitors arranged in parallel. In this case, the same potential difference is applied across both the capacitors. But the plate charges ($\pm Q_1$) on capacitor 1 and the plate charges ($\pm Q_2$) on capacitor 2 are not necessarily the same:

$$Q_1 = C_1 V, \quad Q_2 = C_2 V \quad (2.51)$$

The equivalent capacitor is one with charge

$$Q = Q_1 + Q_2 \quad (2.52)$$

and potential difference V .

$$Q = CV = C_1 V + C_2 V \quad (2.53)$$

The effective capacitance C is, from Eq. (2.53),

$$C = C_1 + C_2 \quad (2.54)$$

The general formula for effective capacitance C for parallel combination of n capacitors follows similarly [Fig. 2.21(b)] :

$$Q = Q_1 + Q_2 + \dots + Q_n \quad (2.55)$$

$$\text{i.e., } CV = C_1 V + C_2 V + \dots + C_n V \quad (2.56)$$

which gives

$$C = C_1 + C_2 + \dots + C_n \quad (2.57)$$

Answer

- (a) In the given network, C_1 , C_2 and C_3 are connected in series. The effective capacitance C' of these three capacitors is given by

$$\frac{1}{C'} = \frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_3}$$

For $C_1 = C_2 = C_3 = 10 \mu\text{F}$, $C' = (10/3) \mu\text{F}$. The network has C' and C_4 connected in parallel. Thus the equivalent capacitance C of the network is

$$C = C' + C_4 = \left(\frac{10}{3} + 10\right) \mu\text{F} = 13.3 \mu\text{F}$$

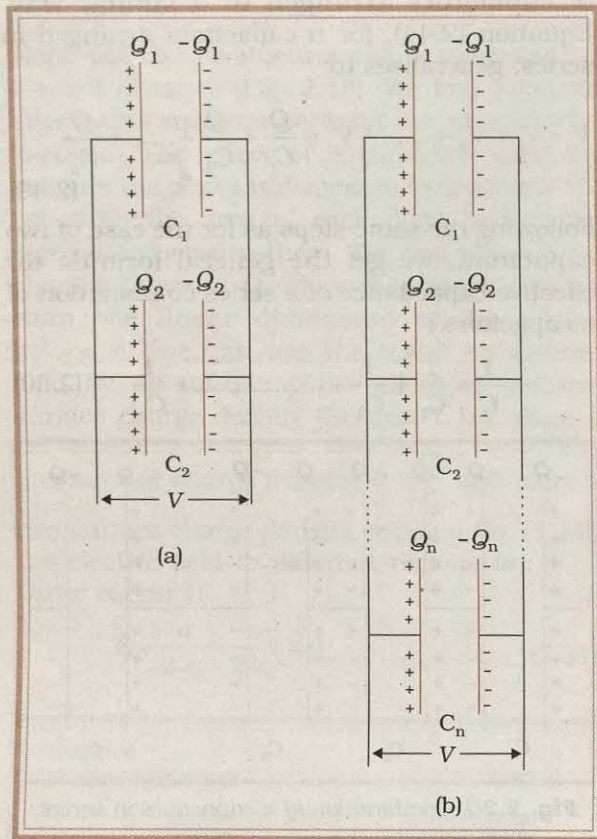


Fig. 2.21 Parallel combination of (a) two capacitors (b) n capacitors.

Example 2.3 A network of four $10 \mu\text{F}$ capacitors is connected to a 500 V supply as shown in Fig. 2.22. Determine the (a) equivalent capacitance of the network and (b) charge on each capacitor. (Note, the 'charge on a capacitor' is the charge on the plate with higher potential, equal and opposite to the charge on the plate with lower potential).

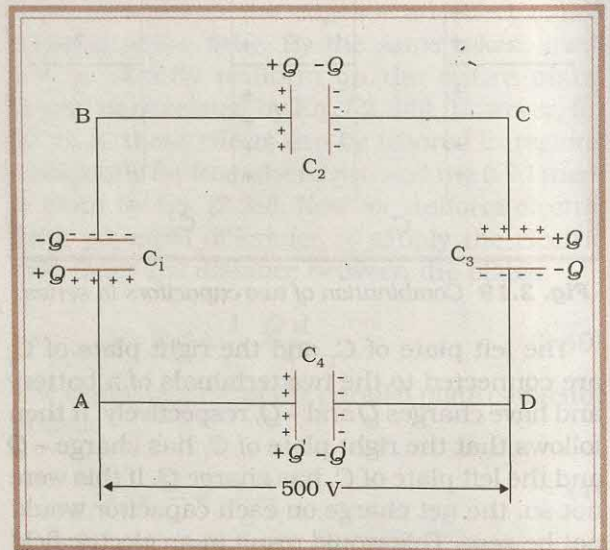


Fig. 2.22 Combination of capacitors in Example 2.3.

- (b) Clearly, from the figure, the charge on each of the capacitors, C_1 , C_2 , and C_3 is the same, say Q . Let the charge on C_4 be Q' . Now, since the potential difference across AB is Q/C_1 , across BC is Q/C_2 , across CD is Q/C_3 , we have

$$\frac{Q}{C_1} + \frac{Q}{C_2} + \frac{Q}{C_3} = 500 \text{ V.}$$

$$\text{Also } Q'/C_4 = 500 \text{ V.}$$

This gives for the given value of the capacitances,

$$Q = 500 \times \frac{10}{3} \mu\text{C} = 1.7 \times 10^{-3} \text{ C and}$$

$$Q' = 10 \times 500 \mu\text{C} = 5.0 \times 10^{-3} \text{ C.}$$

2.13 ENERGY STORED IN A CAPACITOR

A capacitor, as we have seen above, is a system of two conductors with charges Q and $-Q$. To determine the energy stored in this configuration, consider initially two uncharged conductors 1 and 2. Imagine next a process of transferring charge from conductor 2 to conductor 1 bit by bit, so that at the end conductor 1 gets charge Q . By charge conservation, conductor 2 has charge $-Q$ at the end (Fig. 2.23).

In transferring positive charge from conductor 2 to conductor 1, work will be done externally, since at any stage conductor 1 is at a higher potential than conductor 2. To calculate the total work done, we first calculate the work done in a small step involving transfer of an infinitesimal (i.e., vanishingly small) amount of charge. Consider the intermediate situation when the conductors 1 and 2 have charges Q' and $-Q'$ respectively. At this stage, the potential difference V' between conductors 1 and 2 is Q'/C , where C is the capacitance of the system. Next imagine that a small charge $\delta Q'$ is transferred from conductor 2 to 1. Work done in this step ($\delta W'$), resulting in charge Q' on conductor 1 increasing to $Q' + \delta Q'$, is given by

$$\delta W' = V' \delta Q' = \frac{Q'}{C} \delta Q' \quad (2.58)$$

Since $\delta Q'$ can be made as small as we like, Eq. (2.58) can be written as

$$\delta W' = \frac{1}{2C} [(Q' + \delta Q')^2 - Q'^2] \quad (2.59)$$

Eqs. (2.58) and (2.59) are identical because the term second order in $\delta Q'$, i.e. $\delta Q'^2/2C$ is negligible, since $\delta Q'$ is arbitrarily small. The total work done (W) is the sum of the small work ($\delta W'$) over the very large number of steps involved in building the charge Q' from zero to Q :

$$\begin{aligned} W &= \sum_{\text{sum over all steps}} \delta W' \\ &= \sum_{\text{sum over all steps}} \frac{1}{2C} [(Q' + \delta Q')^2 - Q'^2] \end{aligned} \quad (2.60)$$

$$\begin{aligned} &= \frac{1}{2C} [(\delta Q'^2 - 0) + (2\delta Q')^2 - \delta Q'^2] \\ &+ [(3\delta Q')^2 - (2\delta Q')^2] + \dots \\ &+ [Q^2 - (Q - \delta Q)^2] \end{aligned} \quad (2.61)$$

$$= \frac{1}{2C} [Q^2 - 0] = \frac{Q^2}{2C} \quad (2.62)$$

The same result can be obtained directly from Eq. (2.58) by integration:

$$W' = \int_0^Q \frac{Q'}{C} \delta Q' = \frac{1}{C} \frac{Q'^2}{2} \Big|_0^Q = \frac{Q^2}{2C}$$

This is not surprising since integration is nothing but summation of a large number of small terms.

We can write the final result, Eq. (2.62), in different ways:

$$W = \frac{Q^2}{2C} = \frac{1}{2} C V^2 = \frac{1}{2} Q V \quad (2.63)$$

Since electrostatic force is conservative, this work is stored in the form of potential energy of the system. For the same reason, the final result for potential energy [Eq. (2.63)] is independent of the manner in which the charge configuration of the capacitor is built up. When the capacitor discharges, this stored-up energy is released. It is possible to view the potential energy of the capacitor as 'stored' in the electric field between the plates. To see this, consider, for simplicity, a parallel plate capacitor [of area A (of each plate) and separation d between the plates]:

Energy stored in the capacitor

$$= \frac{1}{2} \frac{Q^2}{C} = \frac{(A\sigma)^2}{2} \times \frac{d}{\epsilon_0 A} \quad (2.64)$$

The surface charge density σ is related to the electric field E between the plates:

$$E = \frac{\sigma}{\epsilon_0} \quad (2.65)$$

From Eqs. (2.64) and (2.65), we get

Energy stored in the capacitor

$$= (\frac{1}{2}) \epsilon_0 E^2 \times A d \quad (2.66)$$

Note that $A d$ is the volume of the region between the plates (where alone electric field exists). If we define energy density as energy stored per unit volume of space, Eq. (2.66) shows that Energy density of electric field

$$= (\frac{1}{2}) \epsilon_0 E^2 \quad (2.67)$$

Though we derived Eq. (2.67) for the case of a parallel plate capacitor, the result on energy density of an electric field is, in fact, very general and holds true for electric field due to any configuration of charges.

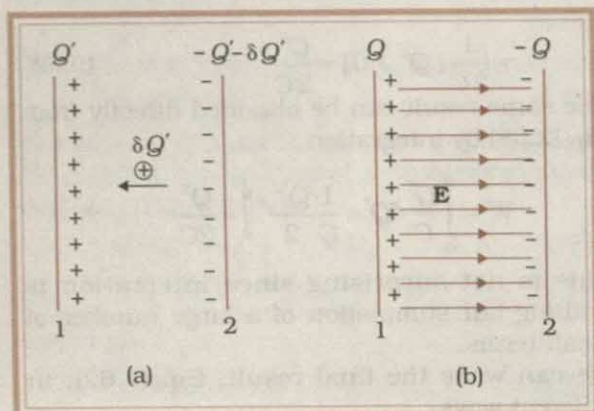


Fig. 2.23 (a) Work done in a small step of building charge on conductor 1 from Q' to $Q' + \delta Q'$. (b) Total work done in charging the capacitor may be viewed as stored in the energy of electric field between the plates.

Example 2.4 (a) A 900 pF capacitor is charged by 100 V battery. How much electrostatic energy is stored by the capacitor? (b) The capacitor is disconnected from the battery and connected to another 900 pF capacitor [Fig. 2.24(b)]. What is the electrostatic energy stored by the system?

Answer

(a) The charge on the capacitor is

$$Q = CV = 900 \times 10^{-12} \text{ F} \times 100 \text{ V} \\ = 9 \times 10^{-8} \text{ C.}$$

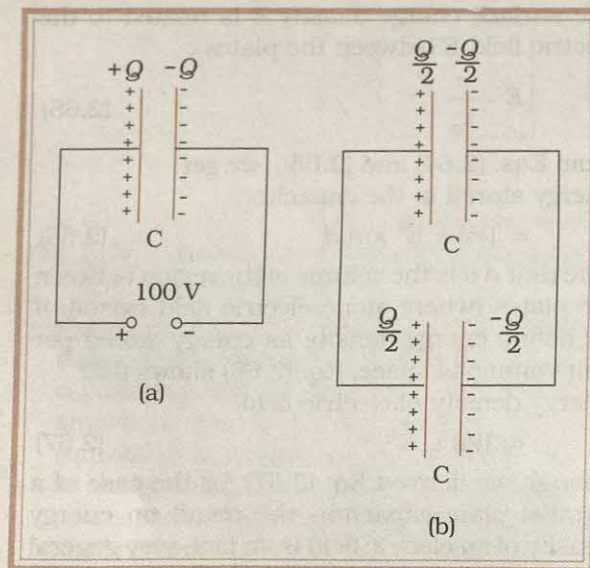


Fig. 2.24 Change in electrostatic energy.

The energy stored by the capacitor is

$$= \frac{1}{2} CV^2 = \frac{1}{2} QV$$

$$= \frac{1}{2} \times 9 \times 10^{-8} \text{ C} \times 100 \text{ V}$$

$$= 4.5 \times 10^{-6} \text{ J}$$

- (b) In the steady situation, the two capacitors have their positive plates at the same potential, and their negative plates at the same potential. Let the common potential difference be V' . The charge on each capacitor is then $Q' = CV'$. By charge conservation, $Q' = Q/2$. This implies $V' = V/2$. The total energy of the system is

$$= 2 \times \frac{1}{2} Q' V' = \frac{1}{4} QV = 2.25 \times 10^{-6} \text{ J}$$

Thus in going from (a) to (b), though no charge is lost, the final energy is only half the initial energy. Where has the remainder of the energy gone?

There is a transient period before the system settles to the situation (b). During this period, a transient current flows from the first capacitor to the second. Energy is lost during this time in the form of heat and electromagnetic radiation.

2.14 DIELECTRICS AND POLARISATION

Dielectrics are non-conducting substances. In contrast to conductors, they have no (or negligible) charge carriers. Recall from Section 2.9 what happens when a conductor is placed in an external electric field. The free charge carriers move and charge-distribution in the conductor adjusts itself in such a way that the electric field due to induced charges opposes the external field. This happens until, in the static situation, the two fields cancel each other and the net electrostatic field in the conductor is zero. In a dielectric, this free movement of charges is not possible. Yet it turns out that the external field induces charges on the surface of the dielectric which produce a field that opposes the external field. Unlike in a conductor, however, the opposing field so induced does not exactly cancel the external field. It only reduces it. The extent of the effect depends on the nature of the dielectric. To understand the effect, we need to look at the charge distribution of a dielectric at the molecular level.

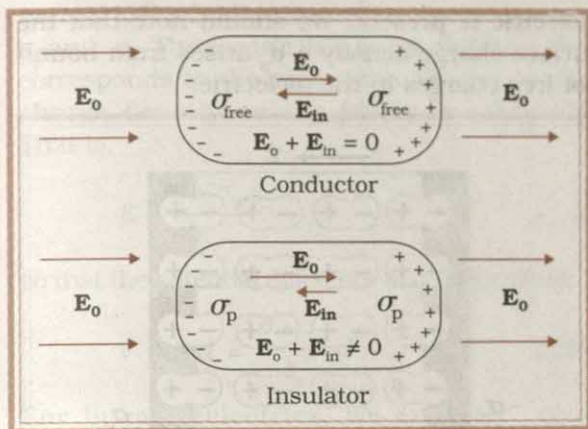


Fig. 2.25 Difference in behaviour of a conductor and a dielectric in an external electric field.

The molecules of a substance may be polar or non-polar. In a non-polar molecule, the centres of positive and negative charges coincide. The molecule then has no permanent (or intrinsic) dipole moment. Examples of non-polar molecules are oxygen (O_2) and hydrogen (H_2) molecules which, because of their symmetry, have no dipole moment. On the other hand, a polar molecule is one in which the centres of positive and negative charges are separated (even when there is no external field). Such molecules have permanent dipole moment. An ionic molecule such as HCl or a molecule of water (H_2O) are examples of polar molecules.

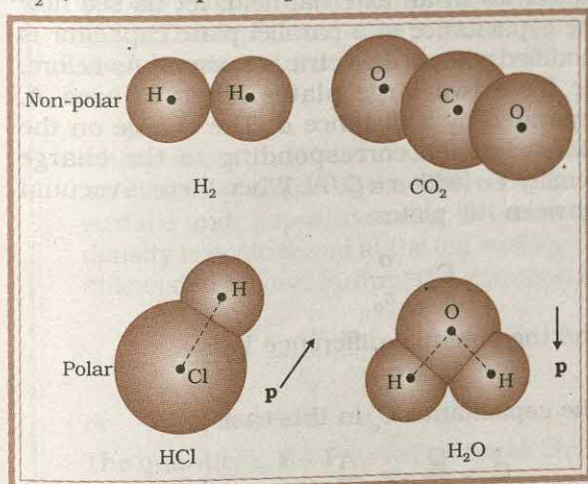


Fig. 2.26 Some examples of polar and non-polar molecules.

In an external electric field, the positive and negative charges of a non-polar molecule are displaced in opposite directions. The displacement stops when the external force on the constituent charges of the molecule is balanced by the restoring force (due to internal fields in the molecule). The non-polar molecule thus develops an induced dipole moment. The dielectric is said to be polarised by the external field. We consider only the simple situation when the induced dipole moment is in the direction of the field and is proportional to the field strength*. The induced dipole moments of different molecules add up giving a net dipole moment of the dielectric in the presence of the external field.

A dielectric with polar molecules also develops a net dipole moment in an external field, but for a different reason. In the absence of any external field, the different permanent dipoles are oriented randomly due to thermal agitation; so the total dipole moment is zero. When an external field is applied, the individual dipole moments tend to align with the field. When summed over all the molecules, there is then a

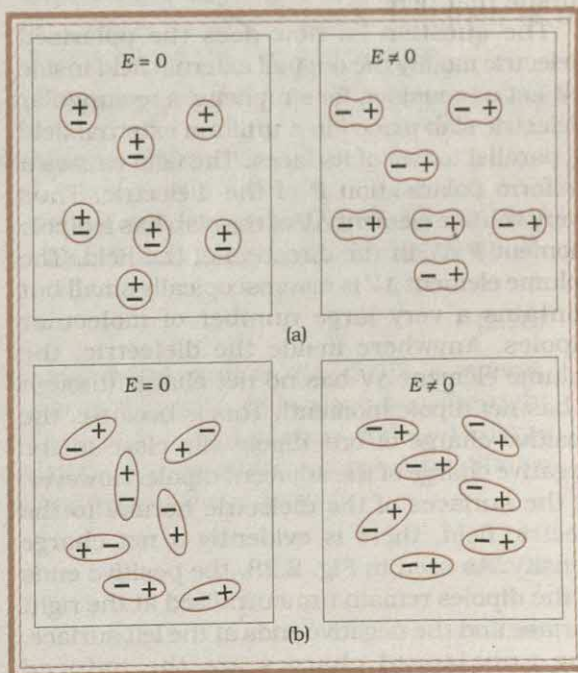


Fig. 2.27 How a dielectric develops a net dipole moment in an external electric field. (a) Non-polar molecules (b) Polar molecules.

* Substances for which this assumption is true are called linear isotropic dielectrics. For a non-isotropic dielectric, polarization may be in a different direction from that of the electric field. For high electric fields, non-linear effects may set in i.e., polarization may not only depend on E but also on higher powers of E . We shall not pursue these matters here.

net dipole moment in the direction of the external field i.e., the dielectric is polarised. The extent of polarisation depends on the relative strength of two mutually opposing factors: the dipole potential energy in the external field tending to align the dipoles with the field and the thermal energy tending to disrupt the alignment. There may be, in addition, the 'induced dipole moment' effect as for non-polar molecules, but generally the alignment effect is more important for polar molecules.

Thus in either case, whether polar or non-polar, a dielectric develops a net dipole moment in the presence of an external field. The dipole moment per unit volume of a substance is called polarisation and is denoted by \mathbf{P} . For linear isotropic dielectrics,

$$\mathbf{P} = \chi_e \mathbf{E} \quad (2.68)$$

where χ_e is a constant characteristic of the dielectric and is known as electric susceptibility of the dielectric medium.

It is possible to relate χ_e to the molecular properties of the substance, but we shall not pursue that here.

The question is: how does the polarised dielectric modify the original external field inside it? Let us consider, for simplicity, a rectangular dielectric slab placed in a uniform external field \mathbf{E}_0 parallel to two of its faces. The field causes a uniform polarisation \mathbf{P} of the dielectric. Thus every volume element ΔV of the slab has a dipole moment $\mathbf{P} \Delta V$ in the direction of the field. The volume element ΔV is macroscopically small but contains a very large number of molecular dipoles. Anywhere inside the dielectric, the volume element ΔV has no net charge (though it has net dipole moment). This is because, the positive charge of one dipole sits close to the negative charge of the adjacent dipole. However, at the surfaces of the dielectric normal to the electric field, there is evidently a net charge density. As seen in Fig. 2.28, the positive ends of the dipoles remain unneutralised at the right surface and the negative ends at the left surface. The unbalanced charges are the induced charges due to the external field.

Thus the polarised dielectric is equivalent to two charged surfaces with induced surface charge densities, say σ_p and $-\sigma_p$. Clearly, the field produced by these surface charges opposes the external field. The total field in the dielectric is, thereby, reduced from the case when no

dielectric is present. We should note that the surface charge density $\pm \sigma_p$ arises from bound not free charges in the dielectric.

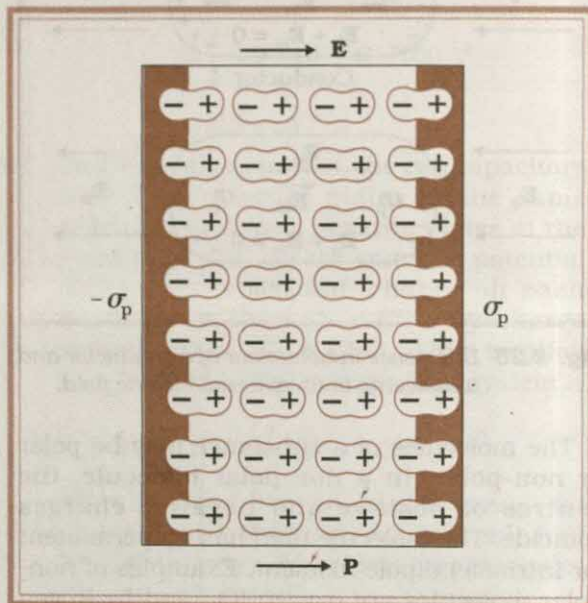


Fig. 2.28 A uniformly polarized dielectric amounts to induced surface charge density, but no volume charge density.

2.15 EFFECT OF DIELECTRIC ON CAPACITANCE

With this understanding of the behaviour of dielectrics in an external field, let us see how the capacitance of a parallel plate capacitor is modified when a dielectric is present. As before, we have two large plates, each of area A , separated by a distance d . The charge on the plates is $\pm Q$, corresponding to the charge density $\pm \sigma$ (with $\sigma = Q/A$). When there is vacuum between the plates,

$$E_0 = \frac{\sigma}{\epsilon_0}$$

and the potential difference V_0 is

$$V_0 = E_0 d$$

The capacitance C_0 in this case is

$$C_0 = \frac{Q}{V_0} = \epsilon_0 \frac{A}{d} \quad (2.69)$$

Consider next a dielectric inserted between the plates fully occupying the intervening region. The dielectric is polarised by the field and, as explained above, the effect is equivalent to two charged sheets (at the surfaces of the dielectric

normal to the field) with surface charge densities σ_p and $-\sigma_p$. The electric field in the dielectric then corresponds to the case when the net surface charge density on the plates is $\pm (\sigma - \sigma_p)$. That is,

$$E = \frac{\sigma - \sigma_p}{\epsilon_0} \quad (2.70)$$

so that the potential difference across the plates is

$$V = E d = \frac{\sigma - \sigma_p}{\epsilon_0} d \quad (2.71)$$

For linear dielectrics, we expect σ_p to be proportional to E_0 i.e., to σ . Thus $\sigma - \sigma_p$ is proportional to σ and we can write :

$$\sigma - \sigma_p = \frac{\sigma}{K} \quad (2.72)$$

where K is a constant characteristic of the dielectric. Clearly, $K > 1$. We then have

$$V = \frac{\sigma d}{\epsilon_0 K} = \frac{Q d}{A \epsilon_0 K} \quad (2.73)$$

The capacitance C , with the dielectric between the plates, is then

$$C = \frac{Q}{V} = \frac{\epsilon_0 K A}{d} \quad (2.74)$$

The product $\epsilon_0 K$ is called the permittivity of the medium and is denoted by ϵ :

$$\epsilon = \epsilon_0 K \quad (2.75)$$

For vacuum $K = 1$ and $\epsilon = \epsilon_0$; ϵ_0 is called the permittivity of vacuum. The dimensionless ratio

$$K = \frac{\epsilon}{\epsilon_0} \quad (2.76)$$

is called the dielectric constant of the substance. As remarked before, from Eq. (2.72), it is clear that K is greater than 1. From Eqs. (2.69) and (2.74),

$$K = \frac{C}{C_0} \quad (2.77)$$

Thus, the dielectric constant of a substance is the factor (>1) by which the capacitance increases from its vacuum value, when the dielectric is inserted fully between the plates of a capacitor. Though we arrived at Eq. (2.77) for the case of a parallel plate capacitor, it holds good for any type of capacitor and can, in fact, be viewed in general as a definition of the dielectric constant of a substance.

The Electric Displacement Vector

We have introduced the notion of dielectric constant and arrived at Eq. (2.77), without giving the explicit relation between the induced charge density σ_p and the polarisation \mathbf{P} . We take without proof the result that

$$\sigma_p = \mathbf{P} \cdot \hat{\mathbf{n}} \quad (2.78)$$

where $\hat{\mathbf{n}}$ is a unit vector along the outward normal to the surface. Eq. (2.78) is general, true for any shape of the dielectric. For the slab in Fig. 2.28, \mathbf{P} is along $\hat{\mathbf{n}}$ at the right surface and opposite to $\hat{\mathbf{n}}$ at the left surface. Thus at the right surface, induced charge density is positive and at the left surface, it is negative, as guessed already in our qualitative discussion before. Putting the equation for electric field in vector form

$$\mathbf{E} \cdot \hat{\mathbf{n}} = \frac{\sigma - \mathbf{P} \cdot \hat{\mathbf{n}}}{\epsilon_0} \quad (2.79)$$

$$\text{or } (\epsilon_0 \mathbf{E} + \mathbf{P}) \cdot \hat{\mathbf{n}} = \sigma \quad (2.80)$$

The quantity $\epsilon_0 \mathbf{E} + \mathbf{P}$ is called the electric displacement vector and is denoted by \mathbf{D} . Thus,

$$\mathbf{D} \cdot \hat{\mathbf{n}} = \sigma, \quad \mathbf{D} = \epsilon_0 \mathbf{E} + \mathbf{P} \quad (2.81)$$

The significance of \mathbf{D} is this : in vacuum, \mathbf{E} is related to the free charge density σ . When a dielectric medium is present, the corresponding role is taken up by \mathbf{D} . For a dielectric medium, it is \mathbf{D} not \mathbf{E} that is directly related to free charge density σ , as seen in Eq. (2.81). Since \mathbf{P} is in the same direction as \mathbf{E} , all the three vectors \mathbf{P} , \mathbf{E} and \mathbf{D} are parallel.

From Eqs. (2.81) and (2.79), the ratio of the magnitudes of \mathbf{D} and \mathbf{E} is

$$\frac{D}{E} = \frac{\sigma \epsilon_0}{\sigma - \sigma_p} = \epsilon_0 K \quad (2.82)$$

Thus

$$\mathbf{D} = \epsilon_0 K \mathbf{E} \quad (2.83)$$

$$\text{and } \mathbf{P} = \mathbf{D} - \epsilon_0 \mathbf{E} = \epsilon_0 (K - 1) \mathbf{E} \quad (2.84)$$

This gives for the electric susceptibility χ_e defined in Eq. (2.68):

$$\chi_e = \epsilon_0 (K - 1) \quad (2.85)$$

Example 2.5 A slab of material of dielectric constant K has the same area as the plates of a parallel-plate capacitor but has a thickness $(3/4)d$, where d is the separation of the plates. How is the capacitance changed when the slab is inserted between the plates?

Answer Let $E_0 = V/d$ be the electric field between the plates when there is no dielectric and the potential difference is V_0 . If the dielectric is now inserted, the electric field in the dielectric will be $E = E_0/K$. The potential difference will then be

$$\begin{aligned} V &= E_0 \left(\frac{1}{4}d \right) + \frac{E_0}{K} \left(\frac{3}{4}d \right) \\ &= E_0 d \left(\frac{1}{4} + \frac{3}{4K} \right) = V_0 \frac{K+3}{4K} \end{aligned}$$

The potential difference decreases by the factor $(K+3)/4K$ while the free charge Q_0 on the plates remains unchanged. The capacitance thus increases :

$$C = \frac{Q_0}{V} = \frac{4K}{K+3} \frac{Q_0}{V_0} = \frac{4K}{K+3} C_0$$

2.16 VAN DE GRAAFF GENERATOR

This is a machine that can build up high voltages of the order of a few million volts. The resulting large electric fields are used to accelerate charged particles (electrons, protons, ions) to high energies needed for experiments to probe the small scale structure of matter. The principle underlying the machine is as follows:

Consider a large spherical conducting shell of radius R with uniform charge density.

The total charge on the shell is Q . We know from Chapter 1 that the field at distance $r \geq R$ is like the field produced by a charge Q at the centre for the shell. The corresponding potential is given by

$$\phi_Q = \frac{Q}{4\pi\epsilon_0 r} \quad r \geq R \quad (2.86)$$

Since the field inside the shell is zero, the potential inside is constant, equal to its value at the surface.

$$\phi_Q = \frac{Q}{4\pi\epsilon_0 R} \quad r \leq R \quad (2.87)$$

Next imagine that a sphere of radius r_0 (with $r_0 < R$) carrying a total charge q (uniformly distributed over it) is introduced in the larger conducting shell and placed at its centre. The potential due to the inner sphere is given by

$$\phi_q = \frac{q}{4\pi\epsilon_0 r} \quad r \geq r_0 \quad (2.88)$$

Using the superposition principle, the potential due to the system (sphere inside the shell) can be found out from Eqs. (2.87) and (2.88) at $r = R$ and $r = r_0$:

$$\phi(R) = \frac{1}{4\pi\epsilon_0} \left(\frac{Q}{R} + \frac{q}{R} \right) \quad (2.89a)$$

$$\phi(r_0) = \frac{1}{4\pi\epsilon_0} \left(\frac{Q}{R} + \frac{q}{r_0} \right) \quad (2.89b)$$

The potential difference between the points $r = r_0$ and $r = R$ is, from Eqs. (2.89) :

$$\phi(r_0) - \phi(R) = \frac{q}{4\pi\epsilon_0} \left(\frac{1}{r_0} - \frac{1}{R} \right) \quad (2.90)$$

Thus for positive q , whatever be the magnitude and sign of Q , the small sphere is at a higher potential than the shell. If an electric contact is provided, charge would flow from the small sphere to the shell. By repeating the process, a large amount of charge can be piled up on the shell, with consequent high potential and field.

Figure 2.29 shows schematically the way this principle is implemented in practice. An insulating column several metres high supports a large spherical conducting shell with radius of a few metres. There are two pulleys, one at the centre of the shell and the other at the ground. A long narrow belt made of insulating material passes over the pulleys. Charge is sprayed on to the belt at the lower pulley by means of a discharge through a metallic brush (with sharp points) connected to a high voltage source. The belt is moved rapidly by a motor driving the lower pulley. The (positive) charge thus transferred continuously upward is removed from the belt, again by a metallic brush connected to the shell. In this way, the shell

builds up a huge voltage. The machine can generate high energy beams in the range of 10 MeV or so.

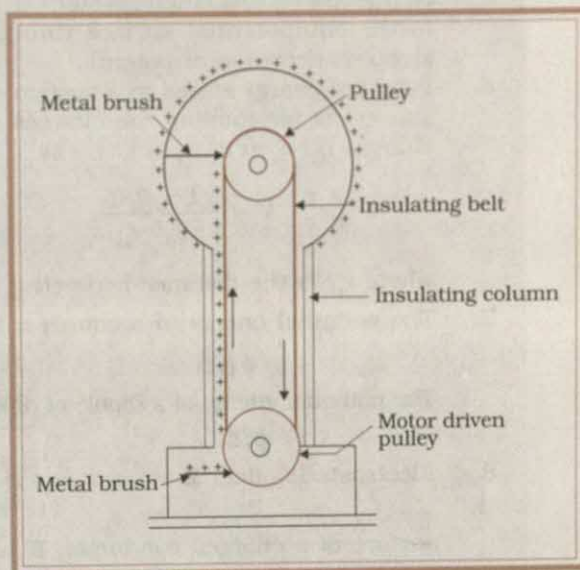


Fig. 2.29 Schematic diagram of a Van de Graaff generator.

SUMMARY

1. Work done by an external force (equal and opposite to the electrostatic force) in bringing a charge q from a point R to a point P is $V_P - V_R$, which is the difference in potential energy of charge q between the final and initial points.
2. Potential at a point is the work done per unit charge (by an external agency) in bringing a charge from infinity to that point. Potential at a point is arbitrary to within an additive constant, since it is the potential difference between two points which is physically significant. If potential at infinity is chosen to be zero, potential at a point with position vector \mathbf{r} due to a point charge Q placed at the origin is given by

$$\phi(\mathbf{r}) = \frac{1}{4\pi\epsilon_0} \frac{Q}{r}$$

3. The electrostatic potential at a point with position vector \mathbf{r} due to a point dipole of dipole moment \mathbf{p} placed at the origin is

$$\phi = \frac{1}{4\pi\epsilon_0} \frac{\mathbf{p} \cdot \hat{\mathbf{r}}}{r^2}$$

The result is true for a dipole (with charges $-q$ and q separated by $2a$) for $r \gg a$.

4. For a charge configuration q_1, q_2, \dots, q_n with position vectors $\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_n$, the potential at a point P is given by the superposition principle

$$\phi = \frac{1}{4\pi\epsilon_0} \left(\frac{q_1}{r_{1P}} + \frac{q_2}{r_{2P}} + \dots + \frac{q_n}{r_{nP}} \right)$$

where r_{1P} is the distance between q_1 and P , and so on.

5. An equipotential surface is a surface over which potential has a constant value. For a point charge, concentric spheres centred at the location of the charge are equipotential surfaces. The electric field \mathbf{E} at a point is perpendicular to the equipotential surface through the point. \mathbf{E} is in the direction of the steepest decrease of potential.
6. Potential energy stored in a system of charges is the work done (by an external agency) in assembling the charges at their locations. Potential energy of two charges q_1, q_2 at $\mathbf{r}_1, \mathbf{r}_2$ is given by

$$V = \frac{1}{4\pi\epsilon_0} \frac{q_1 q_2}{r_{12}}$$

where r_{12} is the distance between q_1 and q_2 .

7. The potential energy of a charge q in an external potential $\phi(\mathbf{r})$ is

$$q \phi(\mathbf{r})$$

The potential energy of a dipole of dipole moment \mathbf{p} in a uniform electric field \mathbf{E} is $-\mathbf{p} \cdot \mathbf{E}$

8. Electrostatic field \mathbf{E} is zero in the interior of a conductor; just outside the

surface of a charged conductor, \mathbf{E} is normal to the surface given by $\mathbf{E} = \frac{\sigma}{\epsilon_0} \hat{\mathbf{n}}$

where $\hat{\mathbf{n}}$ is the unit vector along the outward normal to the surface and σ is the surface charge density. Charges in a conductor can reside only at its surface. Potential is constant within and on the surface of a conductor. In a cavity within a conductor, the electric field is zero.

9. A capacitor is a system of two conductors separated by an insulator. Its capacitance C is defined by $C = Q/V$, where Q and $-Q$ are the charges on the two conductors and V is the potential difference between them. C is determined purely geometrically, by the shapes, sizes and relative positions of the two conductors. The unit of capacitance is farad: $1 \text{ F} = 1 \text{ C V}^{-1}$.
For a parallel plate capacitor (with vacuum between the plates),

$$C = \epsilon_0 A/d$$

where A is the area of each plate and d the separation between them.

10. For capacitors in the series combination, the total capacitance C is given by

$$\frac{1}{C} = \frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_3} + \dots$$

In the parallel combination, the total capacitance C is :

$$C = C_1 + C_2 + C_3 + \dots$$

where C_1, C_2, C_3, \dots are individual capacitances.

11. The energy E stored in a capacitor of capacitance C , with charge Q and voltage V is

$$E = \frac{1}{2} QV = \frac{1}{2} CV^2 = \frac{1}{2} \frac{Q^2}{C}$$

The electrical energy density (energy per unit volume) in a region with electric field is $(1/2) \epsilon_0 E^2$.

12. If the medium between the plates of a capacitor is filled with an insulating substance (dielectric), the electric field due to the charged plates induces a net dipole moment in the dielectric. This effect called polarisation gives rise to

a field in the opposite direction. The net electric field inside the dielectric and hence the potential difference between the plates are thus reduced. Consequently, the capacitance C increases from its value C_0 when there is no medium (vacuum):

$$C = K C_0$$

where K is the dielectric constant of the insulating substance.

13. A Van de Graaff generator consists of a large spherical conducting shell (a few metre in diameter). By means of a moving belt and suitable brushes, charge is continuously transferred to the shell, and potential difference of the order of several million volts is built up, which can be used for accelerating charged particles.

Physical quantity	Symbol	Dimensions	Unit	Remarks
Potential	ϕ or V	$[M L^2 T^{-3} A^{-1}]$	V	Potential <i>difference</i> is physically significant
Capacitance	C	$[M^{-1} L^{-2} T^4 A^2]$	F	
Polarisation	\mathbf{P}	$[L^{-2} AT]$	$C m^{-2}$	Dipole moment per unit volume
Dielectric constant	K	[Dimensionless]		

POINTS TO PONDER

- Electrostatics deals with forces between charges at rest. But if there is a force on a charge, how can it be at rest? Thus, when we are talking of electrostatic force between charges, it should be understood that each charge is being kept at rest by some unspecified force that opposes the net Coulomb force on the charge.
- A capacitor is so configured that it confines the electric field lines within a small region of space. Thus, even though field may have considerable strength, the potential difference between the two conductors of a capacitor is small.
- Electric field is discontinuous across the surface of a spherical charged shell.

It is zero inside and $\frac{\sigma}{\epsilon_0} \hat{n}$ outside. Electric potential is, however, continuous

across the surface, equal to $\frac{q}{4\pi\epsilon_0 R}$ at the surface.

- The torque $\mathbf{p} \times \mathbf{E}$ on a dipole causes it to oscillate about \mathbf{E} . Only if there is a dissipative mechanism, the oscillations are damped and the dipole eventually aligns with \mathbf{E} .
- Potential due to a charge q at its own location is not defined – it is infinite.
- In the expression $q \phi(\mathbf{r})$ for potential energy of a charge q , $\phi(\mathbf{r})$ is the potential due to external charges and not the potential due to q . As seen in point 5, this expression will be ill defined if $\phi(\mathbf{r})$ includes potential due to the charge q itself.
- A cavity inside a conductor is shielded from outside electrical influences. It is worth noting that electrostatic shielding does not work the other way round; that is, if you put charges inside the cavity, the exterior of the conductor is not shielded from the fields by the inside charges.

EXERCISES

- 2.1** Two charges $5 \times 10^{-8} \text{ C}$ and $-3 \times 10^{-8} \text{ C}$ are located 16 cm apart. At what point on the line joining the two charges is the electric potential zero? Take the potential at infinity to be zero.
- 2.2** A regular hexagon of side 10 cm has a charge $5 \mu\text{C}$ at each of its vertices. Calculate the potential at the centre of the hexagon.
- 2.3** (a) Calculate the potential at a point P due to a charge of $4 \times 10^{-7} \text{ C}$ located 9 cm away.
(b) Hence obtain the work done in bringing a charge of $2 \times 10^{-9} \text{ C}$ from infinity to the point P. Does the answer depend on the path along which the charge is brought?
- 2.4** (a) Determine the electrostatic potential energy of a system containing two charges $7 \mu\text{C}$ and $-2 \mu\text{C}$ separated by a distance of 18 cm.
(b) How much work is required to separate the two charges infinitely away from each other?
- 2.5** Two charges $2 \mu\text{C}$ and $-2 \mu\text{C}$ are placed at points A and B 6 cm apart.
(a) Identify an equipotential surface of the system.
(b) What is the direction of the electric field at every point on this surface?
- 2.6** A spherical conductor of radius 12 cm has a charge of $1.6 \times 10^{-7} \text{ C}$ distributed uniformly on its surface. What is the electric field
(a) inside the sphere
(b) just outside the sphere
(c) at a point 18 cm from the centre of the sphere?
- 2.7** A parallel plate capacitor with air between the plates has a capacitance of 8 pF ($1 \text{ pF} = 10^{-12} \text{ F}$). What will be the capacitance if the distance between the plates is reduced by half, and the space between them is filled with a substance of dielectric constant 6?
- 2.8** Three capacitors each of capacitance 9 pF are connected in series.
(a) What is the total capacitance of the combination?
(b) What is the potential difference across each capacitor if the combination is connected to a 120 V supply?
- 2.9** Three capacitors of capacitances 2 pF, 3 pF and 4 pF are connected in parallel.
(a) What is the total capacitance of the combination?
(b) Determine the charge on each capacitor if the combination is connected to a 100 V supply.
- 2.10** In a parallel plate capacitor with air between the plates, each plate has an area of $6 \times 10^{-3} \text{ m}^2$ and the distance between the plates is 3 mm. Calculate the capacitance of the capacitor. If this capacitor is connected to a 100 V supply, what is the charge on each plate of the capacitor?
- 2.11** Explain what would happen if in the capacitor given in Exercise 2.10, a 3 mm thick mica sheet (of dielectric constant = 6) were inserted between the plates,
(a) while the voltage supply remained connected.
(b) after the supply was disconnected.
- 2.12** A 12 pF capacitor is connected to a 50 V battery. How much electrostatic energy is stored in the capacitor?

- 2.13** A 600 pF capacitor is charged by a 200 V supply. It is then disconnected from the supply and is connected to another uncharged 600 pF capacitor. How much electrostatic energy is lost in the process?

ADDITIONAL EXERCISES

- 2.14** A charge of 8 mC is located at the origin. Calculate the work done in taking a small charge of -2×10^{-9} C from a point P (0, 0, 3 cm) to a point Q (0, 4 cm, 0), via a point R (0, 6 cm, 9 cm).
- 2.15** A cube of side b has a charge q at each of its vertices. Determine the potential and electric field due to this charge array at the centre of the cube.
- 2.16** Two tiny spheres carrying charges $1.5 \mu\text{C}$ and $2.5 \mu\text{C}$ are located 30 cm apart. Find the potential and electric field :
 (a) at the mid-point of the line joining the two charges, and
 (b) at a point 10 cm from this midpoint in a plane normal to the line and passing through the mid-point.
- 2.17** A spherical conducting shell of inner radius r_1 and outer radius r_2 has a charge Q .
 (a) A charge q is placed at the centre of the shell. What is the surface charge density on the inner and outer surfaces of the shell?
 (b) Is the electric field inside a cavity (with no charge) zero, even if the shell is not spherical, but has any irregular shape? Explain.
- 2.18** (a) Show that the normal component of electrostatic field has a discontinuity from one side of a charged surface to another given by

$$(\mathbf{E}_2 - \mathbf{E}_1) \cdot \hat{\mathbf{n}} = \frac{\sigma}{\epsilon_0}$$

where $\hat{\mathbf{n}}$ is a unit vector normal to the surface at a point and σ is the surface charge density at that point. (The direction of $\hat{\mathbf{n}}$ is from side 1 to side 2.) Hence show that just outside a conductor, the electric field is $\sigma \hat{\mathbf{n}} / \epsilon_0$.

- (b) Show that the tangential component of electrostatic field is continuous from one side of a charged surface to another. [Hint: For (a), use Gauss's law. For, (b) use the fact that work done by electrostatic field on a closed loop is zero.]
- 2.19** A long charged cylinder of linear charged density λ is surrounded by a hollow co-axial conducting cylinder. What is the electric field in the space between the two cylinders?
- 2.20** In a hydrogen atom, the electron and proton are bound at a distance of about 0.53 \AA :
 (a) Estimate the potential energy of the system in eV, taking the zero of the potential energy at infinite separation of the electron from proton.
 (b) What is the minimum work required to free the electron, given that its kinetic energy in the orbit is half the magnitude of potential energy obtained in (a)?
 (c) What are the answers to (a) and (b) above if the zero of potential energy is taken at 1.06 \AA separation?
- 2.21** If one of the two electrons of a H_2 molecule is removed, we get a hydrogen molecular ion (H_2^+). In the ground state of a (H_2^+), the two protons are

separated by roughly 1.5 \AA , and the electron is roughly 1 \AA from each proton. Determine the potential energy of the system. Specify your choice of the zero of potential energy.

- 2.22** Two charged conducting spheres of radii a and b are connected to each other by a wire. What is the ratio of electric fields at the surfaces of the two spheres? Use the result obtained to explain why charge density on the sharp and pointed ends of a conductor is higher than on its flatter portions.

2.23 Answer carefully:

- A comb run through one's dry hair attracts small bits of paper. Why? What happens if the hair is wet or if it is a rainy day? (Remember, a paper does not conduct electricity.)
- Ordinary rubber is an insulator. But the special rubber tyres of aircrafts are made slightly conducting. Why is this necessary?
- Vehicles carrying inflammable materials usually have metallic ropes touching the ground during motion. Why?
- A bird perches on a bare high power line, and nothing happens to the bird. A man standing on the ground touches the same line and gets a fatal shock. Why?

- 2.24** Two charges $-q$ and $+q$ are located at points $(0, 0, -a)$ and $(0, 0, a)$ respectively,

- What is the electrostatic potential at the points $(0, 0, z)$ and $(x, y, 0)$?
- Obtain the dependence of potential on the distance r of a point from the origin when $r/a \gg 1$.
- How much work is done in moving a small test charge from the point $(5, 0, 0)$ to $(-7, 0, 0)$ along the x -axis? Does the answer change if the path of the test charge between the same points is not along the x -axis?

- 2.25** Figures 2.30(a) and (b) show the field lines of a single positive and negative charge respectively.

- Give the signs of the potential difference : $V_P - V_Q$; $V_B - V_A$

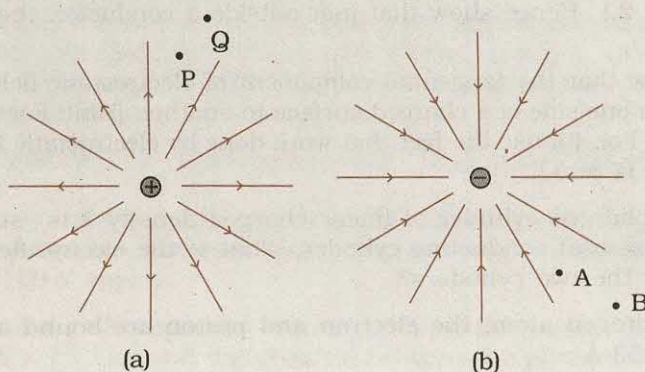


Fig. 2.30

- Give the sign of the potential energy difference of a small negative charge between the points Q and P; A and B.
- Give the sign of the work done by the field in moving a small positive charge from Q to P.
- Give the sign of the work done by an external agency in moving a small negative charge from B to A.
- Does the kinetic energy of a small negative charge increase or decrease in going from B to A?

2.26 A molecule of a substance has permanent electric dipole moment equal to 10^{-29} C m. A mole of this substance is polarised (at low temperature) by applying a strong electrostatic field of magnitude (10^6 V m $^{-1}$). The direction of the field is suddenly changed by an angle of 60° . Estimate the heat released by the substance in aligning its dipoles along the new direction of the field. For simplicity, assume 100% polarisation of the sample.

2.27 Figure 2.31 below shows a charge array known as an 'electric quadrupole'. For a point on the axis of the quadrupole, obtain the dependence of potential on r for $r/a \gg 1$, and contrast your results with that due to an electric dipole, and an electric monopole (i.e., a single charge).

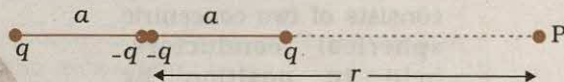


Fig. 2.31

2.28 An electrical technician requires a capacitance of $2 \mu\text{F}$ in a circuit across a potential difference of 1 kV. A large number of $1 \mu\text{F}$ capacitors are available to him each of which can withstand a potential difference of not more than 400 V. Suggest a possible arrangement that requires the minimum number of capacitors.

2.29 What is the area of the plates of a 2 F parallel plate capacitor, given that the separation between the plates is 0.5 cm? [You will realise from your answer why ordinary capacitors are in the range of μF or less. However, electrolytic capacitors do have a much larger capacitance (0.1 F) because of very minute separation between the conductors.]

2.30 Obtain the equivalent capacitance of the network in Fig. 2.32. For a 300 V supply, determine the charge and voltage across each capacitor.

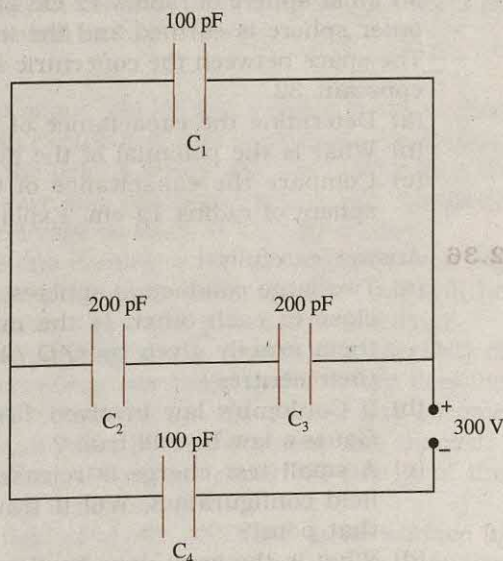


Fig. 2.32

2.31 The plates of a parallel plate capacitor have an area of 90 cm^2 each and are separated by 2.5 mm. The capacitor is charged by connecting it to a 400 V supply.

- How much electrostatic energy is stored by the capacitor?
- View this energy as stored in the electrostatic field between the plates, and obtain the energy per unit volume u . Hence arrive at a relation between u and the magnitude of electric field E between the plates.

- 2.32** A $4\ \mu\text{F}$ capacitor is charged by a $200\ \text{V}$ supply. It is then disconnected from the supply, and is connected to another uncharged $2\ \mu\text{F}$ capacitor. How much electrostatic energy of the first capacitor is lost in the form of heat and electromagnetic radiation?
- 2.33** Show that the force on each plate of a parallel plate capacitor has a magnitude equal to $(1/2)QE$, where Q is the charge on the capacitor, and E is the magnitude of electric field between the plates. Explain the origin of the factor $1/2$.
- 2.34** A spherical capacitor consists of two concentric spherical conductors, held in position by suitable insulating supports. Show that the capacitance of a spherical capacitor is given by

$$C = \frac{4\pi\epsilon_0 r_1 r_2}{r_1 - r_2}$$

where r_1 and r_2 are the radii of outer and inner spheres respectively.

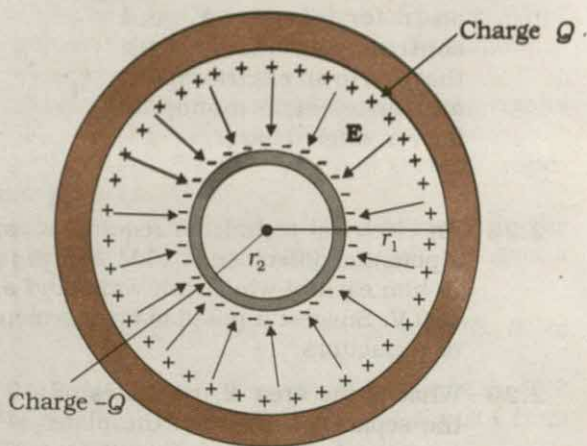


Fig. 2.33

- 2.35** A spherical capacitor has

an inner sphere of radius $12\ \text{cm}$ and an outer sphere of radius $13\ \text{cm}$. The outer sphere is earthed and the inner sphere is given a charge of $2.5\ \mu\text{C}$. The space between the concentric spheres is filled with a liquid of dielectric constant 32 .

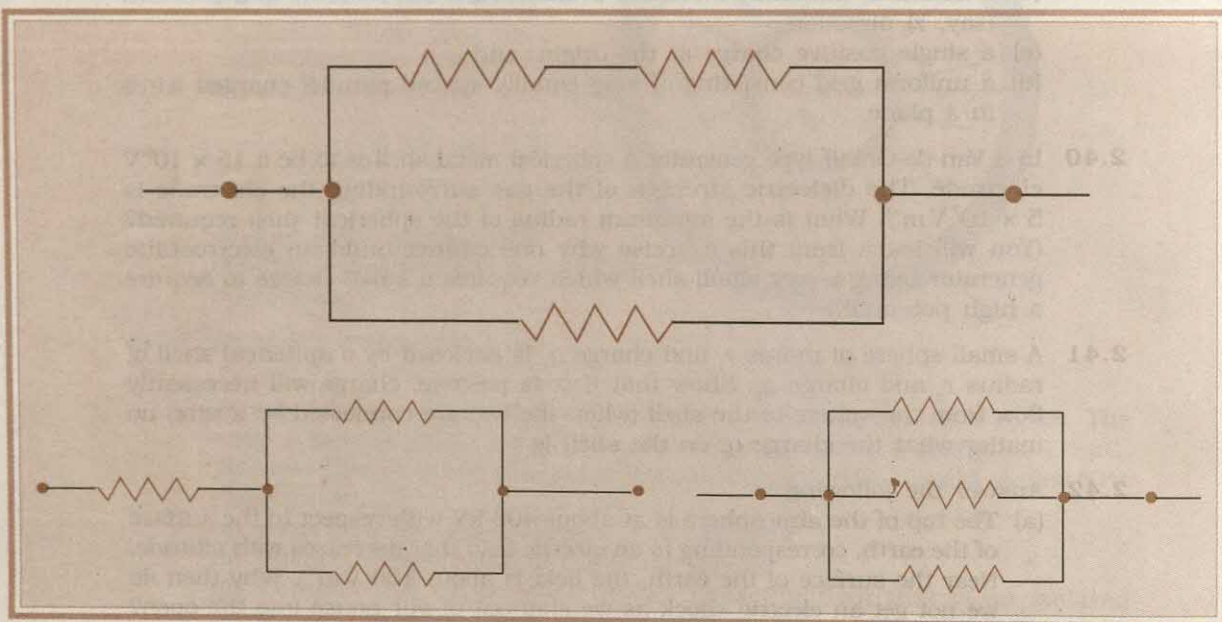
- Determine the capacitance of the capacitor.
 - What is the potential of the inner sphere?
 - Compare the capacitance of this capacitor with that of an isolated sphere of radius $12\ \text{cm}$. Explain why the latter is much smaller.
- 2.36** Answer carefully :
- Two large conducting spheres carrying charges Q_1 and Q_2 are brought close to each other. Is the magnitude of electrostatic force between them exactly given by $Q_1 Q_2 / 4\pi\epsilon_0 r^2$, where r is the distance between their centres?
 - If Coulomb's law involved $1/r^3$ dependence (instead of $1/r^2$), would Gauss's law be still true?
 - A small test charge is released at rest at a point in an electrostatic field configuration. Will it travel along the field line passing through that point?
 - What is the work done by the field of a nucleus in a complete circular orbit of the electron? What if the orbit is elliptical?
 - We know that electric field is discontinuous across the surface of a charged conductor. Is electric potential also discontinuous there?
 - What meaning would you give to the capacity of a single conductor?
 - Guess a possible reason why water has a much greater dielectric constant ($= 80$) than say, mica ($= 6$).
- 2.37** A cylindrical capacitor has two co-axial cylinders of length $15\ \text{cm}$ and radii $1.5\ \text{cm}$ and $1.4\ \text{cm}$. The outer cylinder is earthed and the inner cylinder is

given a charge of $3.5 \mu\text{C}$. Determine the capacitance of the system and the potential of the inner cylinder. Neglect end effects (i.e., bending of field lines at the ends).

- 2.38** A parallel plate capacitor is to be designed with a voltage rating 1 kV, using a material of dielectric constant 3 and dielectric strength about 10^7 Vm^{-1} . (Dielectric strength is the maximum electric field a material can tolerate without breakdown, i.e., without starting to conduct electricity through partial ionisation.) For safety, we should like the field never to exceed, say 10% of the dielectric strength. What minimum area of the plates is required to have a capacitance of 50 pF?
- 2.39** Describe schematically the equipotential surfaces corresponding to
- a constant electric field in the z -direction,
 - a field that uniformly increases in magnitude but remains in a constant (say, z) direction,
 - a single positive charge at the origin, and
 - a uniform grid consisting of long equally spaced parallel charged wires in a plane.
- 2.40** In a Van de Graaff type generator a spherical metal shell is to be a $15 \times 10^6 \text{ V}$ electrode. The dielectric strength of the gas surrounding the electrode is $5 \times 10^7 \text{ Vm}^{-1}$. What is the minimum radius of the spherical shell required? (You will learn from this exercise why one cannot build an electrostatic generator using a very small shell which requires a small charge to acquire a high potential.)
- 2.41** A small sphere of radius r_1 and charge q_1 is enclosed by a spherical shell of radius r_2 and charge q_2 . Show that if q_1 is positive, charge will necessarily flow from the sphere to the shell (when the two are connected by a wire) no matter what the charge q_2 on the shell is.
- 2.42** Answer the following :
- The top of the atmosphere is at about 400 kV with respect to the surface of the earth, corresponding to an electric field that decreases with altitude. Near the surface of the earth, the field is about 100 Vm^{-1} . Why then do we not get an electric shock as we step out of our house into the open? (Assume the house to be a steel cage so there is no field inside!)
 - A man fixes outside his house one evening a two metre high insulating slab carrying on its top a large aluminium sheet of area 1 m^2 . Will he get an electric shock if he touches the metal sheet next morning?
 - The discharging current in the atmosphere due to the small conductivity of air is known to be 1800 A on an average over the globe. Why then does the atmosphere not discharge itself completely in due course and become electrically neutral? In other words, what keeps the atmosphere charged?
 - What are the forms of energy into which the electrical energy of the atmosphere is dissipated during a lightning?
- (Hint: The earth has an electric field of about 100 Vm^{-1} at its surface in the downward direction, corresponding to a surface charge density $= -10^{-9} \text{ C m}^{-2}$. Due to the slight conductivity of the atmosphere up to about 50 km (beyond which it is good conductor), about $+ 1800 \text{ C}$ is pumped every second into the earth as a whole. The earth, however, does not get discharged since thunderstorms and lightning occurring continually all over the globe pump an equal amount of negative charge on the earth.)

CHAPTER THREE

CURRENT ELECTRICITY



3.1 INTRODUCTION

Just as the flow of water in a river constitutes a current of water, flow of charge, i.e., charges in motion constitute an electric current. We have seen in Chapter 1 that when an isolated conductor is placed in an electric field, the charges in the conductor rearrange themselves so that the interior of the conductor has no net electric field. The rearrangement of charges constitutes a current of short duration, called a transient current. The transient current ceases the moment the net electric field in the conductor becomes zero. Another example of a transient current is when we see lightning, which is the flow of electric charge between two clouds or from a cloud to the earth. But we are also familiar with continuous steady currents, such as in a battery torch. The current goes on as long as the torch is on (if the battery has not run out of use!) producing heat in the bulb and light.

How does one maintain such a steady current? In this Chapter, we look into the essential requirements for maintaining a steady current — a source of emf, as it is called. We also introduce the concept of resistance and attempt to understand it by a qualitative microscopic picture. Finally, we describe the basic laws (Kirchhoff's rules) governing electric currents and their applications.



Georg Simon Ohm (1787-1854)

German physicist, professor at Munich. Ohm was led to his law by an analogy between the conduction of heat: the electric field is analogous to the temperature gradient, and the electric current is analogous to the heat flow.



Gustav Robert Kirchhoff (1824-1887)

German physicist, professor at Heidelberg and at Berlin. Mainly known for his development of spectroscopy, he also made many important contributions to mathematical physics, among them, his first and second rules for circuits.

3.2 ELECTRIC CURRENT

You may have experienced the strong water current of the Ganges at a place like Haridwar. At some other place, the current may not be so strong. How to quantify the water current of a river? We can hold a ring normal to the direction of water flow and measure the amount of water that flows through the ring per second. The same idea is used to define electric current in a conductor. Electric current across an area held perpendicular to the direction of flow of charge is defined to be the amount of charge flowing across the area per unit time. If charge ΔQ passes through the area in time t to $t + \Delta t$, the current I at time t is defined by

$$I = \lim_{\Delta t \rightarrow 0} \frac{\Delta Q}{\Delta t} \quad (3.1)$$

If the current is steady, i.e., it does not change with time, the same equation means

$$I = \frac{Q}{t} \quad (3.2)$$

where Q is the charge that flows across the area in time t . The unit of current is ampere. Ampere (A) is the S.I. base unit, defined in terms of its magnetic effect (Chapter 5). Smaller currents are more conveniently expressed in milliamperes ($1\text{mA} = 10^{-3}\text{A}$) or microamperes ($1\mu\text{A} = 10^{-6}\text{A}$).

3.3 ELECTROMOTIVE FORCE (EMF) AND VOLTAGE

How to maintain a steady electric current in a conductor? Let us go back to the example of water flow. Suppose in a horizontal tube, we wish to maintain a steady water flow from A to B (Fig. 3.1). For this we will need to maintain a steady pressure difference between A and B. Clearly, to maintain a steady flow, water flowing out of B will need to be pumped back to the higher tank and allowed to fall back to A. Thus, an external source of energy (the pump P) is necessary to maintain a steady flow in the horizontal tube. An isolated horizontal tube will not serve the purpose. It will need to be part of a closed circuit that includes the external agency (the pump). Water flows spontaneously from higher to lower pressure. The pump is meant to do the opposite — take water from lower pressure to higher pressure.

For this, the pump will need to work at a steady rate.

A steady electric current in a conductor is maintained in an analogous way. In a conductor, positive charge will flow from higher potential (A) to lower potential (B), i.e., in the direction of the electric field. To maintain a steady electric current, the conductor cannot be isolated; it must be part of a closed circuit that includes an external agency or device (Fig. 3.2). This device is required to transport the positive charge from B back to A, i.e., from lower to higher potential and thus maintain the potential difference between A and B. The external device will need to do work for transporting positive charge from lower to higher potential. Such a device is the source of what is known as *electromotive force* abbreviated as *emf*. It is the analogue of the pump in the water flow circuit. The electrochemical cell and thermocouples (Chapter 4) and the electric generator (Chapter 7) are examples of sources of emf.

The external source, as said above, does work on taking a positive charge from lower to the

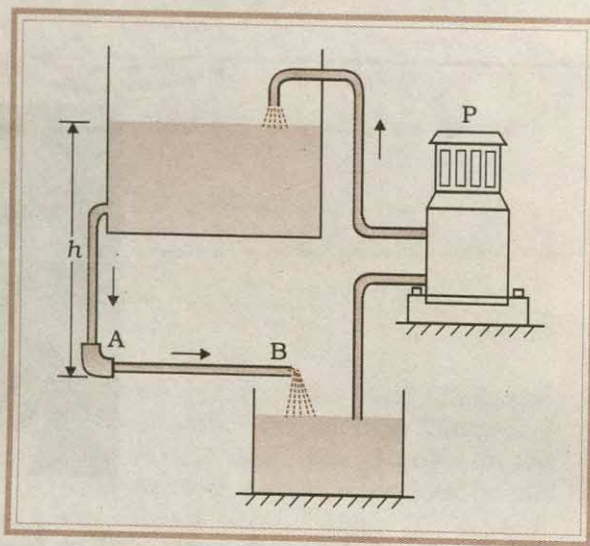


Fig. 3.1 To maintain steady water flow from higher pressure (A) to lower pressure (B), an external device must do work at a steady rate to pump the water from B back to A. Water flows continuously out of a horizontal tube AB at a steady rate if a pressure head (pressure difference between its two ends) is maintained by the tank at a height h , and if a pump P pumps the water back to the tank at the same rate as it flows out at B.

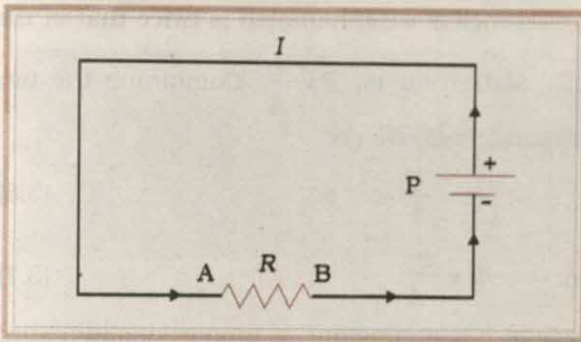


Fig. 3.2 To maintain steady current in an electrical circuit (positive charge flowing from higher potential (A) to lower potential (B)), some device (a source of emf) must do work at a steady rate to take positive charge from lower potential to higher potential. A charge flow circuit is analogous to the water flow circuit of Fig. 3.1. A steady electric current flows across the resistor AB if a source P of emf (e.g., a cell) is connected across it.

higher potential. This is just the opposite to what an electrostatic field does (remember, electric field takes a positive charge from higher to lower potential). A natural way of characterising the external source of energy is in terms of the work that it needs to do per unit positive charge in transporting it from lower to higher potential. This is known as the electromotive force or emf of the device, denoted by ε . The term 'electromotive force' is a misnomer, since it is not a force; rather it is work per unit charge. Perhaps for that reason, nowadays, the abbreviation emf is more commonly used than its expanded version!

In a closed circuit with steady current I , work done by the source of emf per unit charge is used up in two ways: (i) the charge 'falls' through under the electric field from higher to lower potential (A to B in Fig. 3.2). The work done by the field per unit charge is V , where V is the potential difference between the two ends of the conductor A and B; (ii) some work W' is used up in the source itself in transporting the charge. Consequently, by energy conservation,

$$\varepsilon = V + W' \quad (3.3)$$

The important thing to note is that in Eq. (3.3) V is the potential difference across the ends of the external circuit or between the positive and negative terminals of the battery in a closed circuit. In an open circuit, i.e., when the positive

and negative terminals of the battery are not connected by an external conductor, the source of emf does the same work ε per unit charge in establishing potential difference, but in this case, no work is used up in the source, since there is no current in the circuit. That is, in open circuit, $W' = 0$; therefore

$$\varepsilon = V_{\text{open}} \quad (3.4)$$

The emf of a source is thus the potential difference between its two terminals in open circuit.

In introducing the notion of emf above, we considered the current as the flow of positive charge. This is the conventional direction of the current. But actually, we know that it is the negative charge (electrons) that flows in a metallic conductor (while in an electrolyte, both positive and negative ions move). To modify the discussion above for the actual electronic flow, we note that electrons in the external circuit flow from lower to higher potential (against the direction of electric field) and the source of emf needs to do work in taking negative charge from higher to lower potential or taking positive charge from lower to higher potential.

3.4 RESISTANCE AND RESISTIVITY

3.4.1 Ohm's Law

Consider a closed circuit having a source of emf and a conductor in the external circuit. Let the voltage drop across the ends of the conductor be V and let I be the steady current flowing through the conductor. The quantity V/I is a measure of the resistance offered by the conductor for the steady flow of charge through it. The more the resistance, the less is the current I for a given voltage difference, V .

For many substances, it is found experimentally that the ratio V/I is a constant, independent of V or I . The constant of proportionality is called the **resistance** of the substance and is denoted by R .

$$\frac{V}{I} = R \text{ or } V = IR \quad (3.5)$$

Equation (3.5) expresses Ohm's law after G.S. Ohm, who first discovered it in 1828. The constancy of R implies that V and I are linearly related – a graph between measured values of V and I is a straight line. The unit of resistance is ohm ($1 \Omega = 1 \text{ VA}^{-1}$). Ohm's law is only an empirical

law that holds approximately for many substances over certain ranges of V and I . We discuss departures from Ohm's law in a later section.

3.4.2 Resistivity

The resistance of a resistor (an element in a circuit with some resistance R) depends on its geometrical factors (length, cross-sectional area) as also on the nature of the substance of which the resistor is made. It is convenient to separate out the 'size' factors from the resistance R so that we can define a quantity that is characteristic of the material and is independent of the size or shape. Consider a rectangular slab of length l and area of cross section A . For a fixed current I , if the length of the slab is doubled, the potential drop across the slab also doubles. (It is electric field that drives the current in the conductor and potential difference is electric field times the distance). This means that resistance of the slabs doubles with the doubling of its

resistance of each half-slab is twice that of the full slab. That is, $R \propto \frac{l}{A}$. Combining the two dependences, we get

$$R \propto \frac{l}{A} \quad (3.6)$$

$$\text{or} \quad R = \frac{\rho l}{A} \quad (3.7)$$

where ρ is a constant of proportionality called *resistivity*. It depends only on the nature of the material of the resistor and its physical conditions such as temperature and pressure. The unit of resistivity is ohm m (Ω m). The inverse of ρ is called *conductivity*, and is denoted by σ . The unit of σ is $(\Omega\text{m})^{-1}$ or mho m^{-1} or siemen m^{-1} .

A perfect conductor would have zero resistivity and a perfect insulator would have infinite resistivity. Though these are ideal limits, the electrical resistivity of substances has a very wide range (Table 3.1). Metals have low resistivities in the range of 10^{-8} Ω m to 10^{-6} Ω m, while insulators like glass or rubber have resistivity, some 10^{18} times (or even more) greater. Generally, good electrical conductors like metals are also good conductors of heat, while insulators like ceramic or plastic materials are also poor thermal conductors.

The semiconductors form a class intermediate between the conductors and insulators. They are important not primarily because of their resistivities but because of the way they are affected by temperature and small amount of impurities.

Depending on the requirement, we need different kinds of resistors and materials. Transmission of electric power without appreciable loss requires low resistance conductors such as aluminium or copper wires. If an electric current is to be blocked between two points, a good insulator like mica, bakelite, etc., should be used. For moderately high resistances in the range of $\text{k}\Omega$, resistors made of carbon or some semiconducting material are often used.

Commercially produced resistors are of two major varieties: wire-bound resistors and carbon resistors. Wire-bound resistors are made by winding the wires of an alloy, viz., manganin,

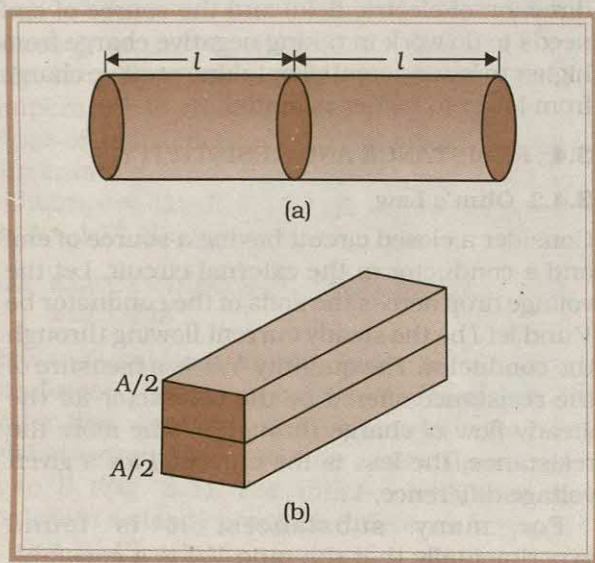


Fig. 3.3 Illustrating the relation $R = \frac{\rho l}{A}$ for a rectangular slab of length l and area of cross-section A .

length. That is, $R \propto l$. Next, imagine the slab as being made of two parallel slabs, each of area $\frac{A}{2}$. If for a given voltage V , the current I flows across the full slab, it is clear that through each half-slab, the current flowing is $\frac{I}{2}$. Thus, the

Table 3.1 Resistivities of some materials

Material	Resistivity, ρ ($\Omega \text{ m}$) at 0°C	Temperature coefficient of resistivity, α $(^\circ\text{C})^{-1}$ $\frac{1}{\rho} \left(\frac{d\rho}{dT} \right)$ at 0°C	Number of outer (valence) electrons per unit cell
A. Conductors			
Silver	1.6×10^{-8}	0.0041	1
Copper	1.7×10^{-8}	0.0068	1
Aluminium	2.7×10^{-8}	0.0043	3
Tungsten	5.6×10^{-8}	0.0045	6
Iron	10×10^{-8}	0.0065	8
Platinum	11×10^{-8}	0.0039	10
Mercury	98×10^{-8}	0.0009	2
Nichrome (alloy of Ni, Fe, Cr)	$\sim 100 \times 10^{-8}$	0.0004	2
Manganin (alloy)	48×10^{-8}	0.002×10^{-3}	
B. Semiconductors			
Carbon (graphite)	3.5×10^{-5}	- 0.0005	4
Germanium	0.46	- 0.05	4
Silicon	2300	- 0.07	4
C. Insulators			
Pure Water	2.5×10^5		
Glass	$10^{10} - 10^{14}$		
Hard Rubber	$10^{13} - 10^{16}$		
NaCl	$\sim 10^{14}$		8
Fused Quartz	$\sim 10^{16}$		

constantan, nichrome on some suitable base. The main advantages of carbon resistors are compactness and low cost.

3.4.3 Colour Code for Resistors

A colour code is used to indicate the resistance value and its percentage accuracy.

The resistor has a set of co-axial coloured rings on it with their significance as indicated in Table 3.2.

The first two bands from the end indicate the first two significant figures of the resistance in ohms. The third band indicates the decimal multiplier and the last band stands for the tolerance or possible variation in percent about the indicated value. If the fourth band is absent, it implies that the tolerance is $\pm 20\%$ (Fig. 3.4).

Table 3.2 Resistor Colour Codes

Colour	Number	Multiplier	Tolerance (%)
Black	0	1	
Brown	1	10^1	
Red	2	10^2	
Orange	3	10^3	
Yellow	4	10^4	
Green	5	10^5	
Blue	6	10^6	
Violet	7	10^7	
Gray	8	10^8	
White	9	10^9	
Gold		10^{-1}	5
Silver		10^{-2}	10
No colour			20

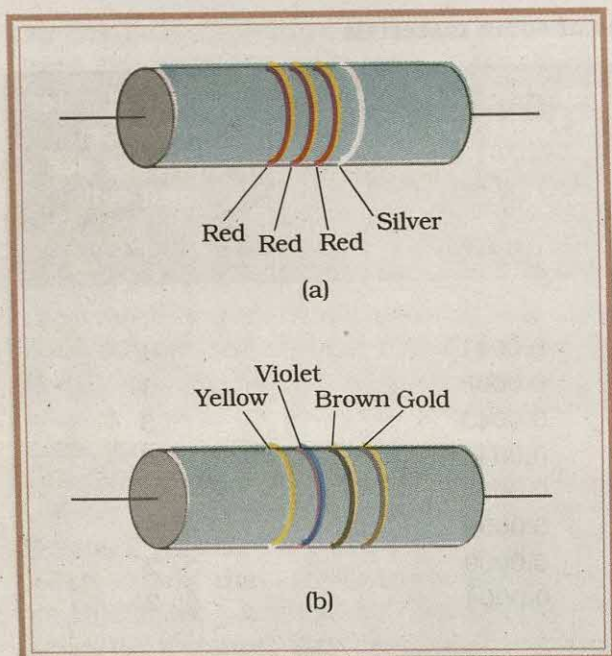


Fig. 3.4 Colour coded resistors

(a) $(22 \times 10^2 \Omega) + 10\%$. (b) $(47 \times 10 \Omega) + 5\%$.

3.5 ORIGIN OF RESISTIVITY

The motion of charge carriers (electrons) in a conductor is very different from that of charges in empty space. In the latter case, under an external electric field, the charge carriers would accelerate. In a conductor, on the other hand, when the current is steady, the charge carriers move with a certain average drift velocity, i.e., on the whole there is no net acceleration. How does this come about? A detailed microscopic theory is involved, but the rough picture is as follows.

At any temperature, the electrons in a metal have a certain distribution of velocities. When there is no external field, all directions are equally likely, and there is no overall drift. In the presence of an external field, each electron

experiences an acceleration of $\frac{eE}{m}$ opposite to the field direction. But this acceleration is momentary, since electrons are continually making random collisions with vibrating atoms or ions or other electrons of the metal. After a collision, each electron makes a fresh start,

accelerates only to be deflected randomly again (Fig. 3.5). If τ is the average time* between two collisions, the average drift speed of the electron is given by

$$v_d = \frac{|e|E}{m} \tau \quad (3.8)$$

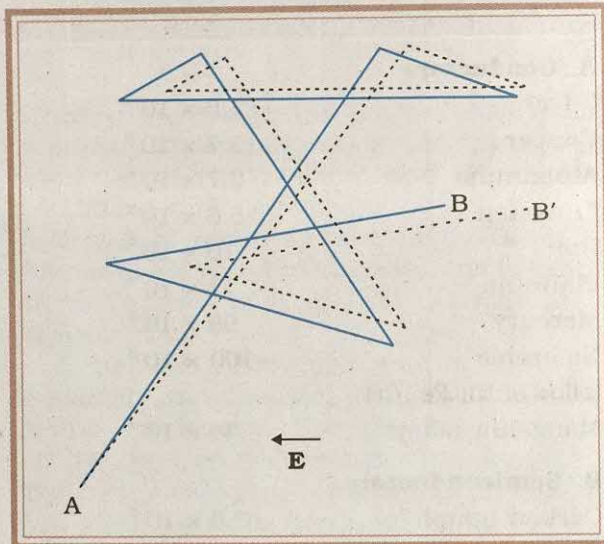


Fig. 3.5 A schematic picture of an electron moving from a point A to another point B through repeated collisions, and straight line travel between collisions (full lines). If an electric field is applied as shown, the electron ends up at point B' (dotted lines). A slight drift in a direction opposite the electric field is visible.

We next define a physical quantity called current density vector, denoted by \mathbf{j} . The direction of \mathbf{j} is the direction of flow of positive charge (or opposite to the direction of drift of electrons in a metal). The magnitude of \mathbf{j} is the amount of charge flowing per unit cross sectional area per second. Thus, if $\delta\mathbf{S}$ is an area element, $\mathbf{j} \cdot \delta\mathbf{S}$ is the amount of charge flowing across the area element per second. If \mathbf{A} is taken to be the cross-sectional area of a wire (with the direction of \mathbf{A} along the conventional current), $\mathbf{j} \cdot \mathbf{A}$ is nothing but the current through the wire. In this case, \mathbf{j} is parallel to \mathbf{A} , so

$$I = j A \quad (3.9)$$

If the drift speed of electrons is v_d , the amount of

*The average time between collisions τ is called the relaxation time because it is a measure of time for the system to relax back to thermal equilibrium through collision.

charge flowing across a unit cross-sectional area in unit time is contained in a cylinder of base of unit area and height v_d , i.e., in a volume $1 \times v_d = v_d$ (Fig. 3.6). If n is the number density of electrons in the metal, i.e., the number of electrons per unit volume, the total magnitude of charge contained in the cylinder of volume v_d is $ne v_d$. Therefore,

$$j = ne v_d \quad (3.10)$$

$$\text{and } I = ne v_d A \quad (3.11)$$

Here, e is the magnitude of electronic charge.

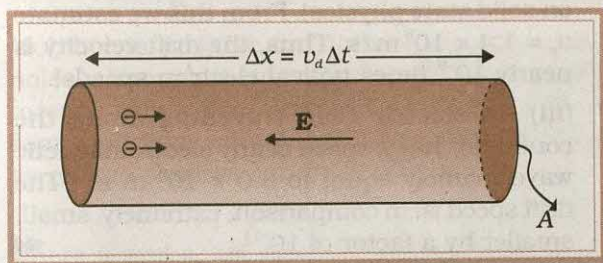


Fig. 3.6 Current in a metallic conductor. The magnitude of current density in a metal is the magnitude of charge contained in a cylinder of unit area and height v_d .

Note that in vector form, Eq. (3.10) can be written as

$$\mathbf{j} = nq \mathbf{v}_d \quad (3.12)$$

where q is the charge of the charge carrier and \mathbf{v}_d the average drift velocity. Equation (3.12) is correct for both signs of q . If $q > 0$, \mathbf{v}_d is in the direction of electric field \mathbf{E} and \mathbf{j} is in the direction of \mathbf{E} . If $q < 0$ ($q = -e$ for electrons), as it is in metallic conductor, \mathbf{v}_d is opposite to \mathbf{E} , and $\mathbf{j} = -ne \mathbf{v}_d$ continues to be in the direction of \mathbf{E} . Substituting for the case of electrons, the vector form of Eq. (3.8)

$$\mathbf{v}_d = -\frac{e\mathbf{E}}{m}\tau$$

in Eq. (3.12), we get

$$\mathbf{j} = (-ne) \times \left(-\frac{e\mathbf{E}}{m}\tau \right)$$

$$\text{i.e., } \mathbf{j} = \frac{ne^2}{m}\tau \mathbf{E} \quad (3.13)$$

Using Eq. (3.9), we may write

$$I = \frac{ne^2}{m}\tau EA$$

Now, if the length of the conductor is l across which there is a potential difference V ,

$$E = \frac{V}{l} \quad (3.14)$$

so that

$$I = \frac{ne^2}{m}\tau \frac{A}{l} V \quad (3.15)$$

$$\text{or } V = \frac{m}{ne^2\tau} \frac{l}{A} I = RI \quad (3.16)$$

We have thus arrived at the Ohm's law on the basis of a simple microscopic picture. From Eq. (3.16), the resistance R can be identified

$$R = \frac{m}{ne^2\tau} \frac{l}{A} \quad (3.17)$$

Using Eq. (3.7), the resistivity ρ of the material is given by

$$\rho = \frac{m}{ne^2\tau} \quad (3.18)$$

We can then rewrite Eq. (3.13) as

$$\mathbf{j} = \frac{\mathbf{E}}{\rho} = \sigma \mathbf{E} \quad (3.19)$$

We see from above that Eqs. (3.19) and (3.16) are quite equivalent. This is why both equations express Ohm's law. The more familiar form $V = IR$ is in terms of directly measurable quantities like V and I , while the form $\mathbf{j} = \sigma \mathbf{E}$ relates the basic vector quantities of the problem, namely, current density vector and the electric field vector.

From the observed value of resistivity of a metal, we can estimate the relaxation time τ for the metal, using Eq. (3.18). For copper, at room temperature, $\rho = 1.7 \times 10^{-8} \Omega \text{ m}$. From the density of copper, one can estimate the number density of electrons to be $8.5 \times 10^{22} \text{ m}^{-3}$. Using the known values of m and e , we get $\tau = 2 \times 10^{-14} \text{ s}$ which agrees with the estimates of τ by other methods. In Example 3.1, the drift speed of electrons for a typical copper wire carrying a current of 1.5 A is calculated. The result on drift speed ($\sim 1 \text{ mm s}^{-1}$) may surprise you. If the drift speed is so small, how is it that an electric bulb lights up as soon as we turn the switch on? We should beware of the misconception that a current starts when an electron from one terminal of the battery reaches the other

terminal! What is needed is that an electric field be established in each region of the conductor which drifts the electrons in that region. That is, the current starts when electrons in different parts of the wire experience electric field and start to drift. The establishment of electric field takes place nearly with the speed of light. This is why the effect of switching on the current is nearly instantaneous.

Example 3.1 (a) Estimate the average drift speed of conduction electrons in a copper wire of cross-sectional area $1.0 \times 10^{-7} \text{ m}^2$ carrying a current of 1.5 A. Assume that each copper atom contributes roughly one conduction electron. The density of copper is $9.0 \times 10^3 \text{ kg/m}^3$, and its atomic mass is 63.5 u. (b) Compare the drift speed obtained above with (i) thermal speeds of copper atoms at ordinary temperatures, (ii) speeds of electrons carrying the current and (iii) speed of propagation of electric field along the conductor which causes the drift motion.

Answer

(a) The direction of drift velocity of conduction electrons is opposite to the electric field direction, i.e., electrons drift in the direction of increasing potential. The drift speed v_d is given by

$$v_d = (I/neA).$$

Now, $e = 1.6 \times 10^{-19} \text{ C}$, $A = 1.0 \times 10^{-7} \text{ m}^2$, $I = 1.5 \text{ A}$. The density of conduction electrons, n is equal to the number of atoms per cubic metre (assuming one conduction electron per Cu atom as is reasonable from its valence electron count of one). A cubic metre of copper has a mass of $9.0 \times 10^3 \text{ kg}$. Since 6.0×10^{23} copper atoms have a mass of 63.5 g,

$$n = \frac{6.0 \times 10^{23}}{63.5} \times 9.0 \times 10^6$$

$$= 8.5 \times 10^{28} \text{ m}^{-3}$$

which gives

$$v_d = \frac{1.5}{8.5 \times 10^{28} \times 1.6 \times 10^{-19} \times 1.0 \times 10^{-7}}$$

$$= 1.1 \times 10^{-3} \text{ m s}^{-1}$$

(b) (i) At a temperature T , the thermal speed of a copper atom of mass M is typically of the order

of $\sqrt{k_B T / M}$, where k_B is the Boltzmann constant. For copper at 300 K, this is about $4 \times 10^4 \text{ m/s}$. This figure indicates the random vibrational speeds of copper atoms in a conductor. Note that the drift speed of electrons is much smaller; about 10^{-7} times the typical thermal speed at ordinary temperatures.

(ii) The maximum kinetic energy $(1/2)mv_F^2$ of electron in copper corresponds to a temperature $T_0 \approx 10^5 \text{ K}$: $(1/2)mv_F^2 = k_B T_0$ with $T_0 \approx 10^5 \text{ K}$. (This information is from a book on solid state physics). From this we estimate $v_F = 1.4 \times 10^6 \text{ m/s}$. Thus, the drift velocity is nearly 10^{-9} times typical electron speeds!

(iii) An electric field travelling along the conductor has a speed of any electromagnetic wave, namely equal to $3.0 \times 10^8 \text{ m s}^{-1}$. The drift speed is, in comparison, extremely small; smaller by a factor of 10^{-11} .

Example 3.2 A potential difference of 100 V is applied to the ends of a copper wire one metre long. Calculate the average drift velocity of the electrons. Compare it with thermal velocity at 27°C . (Use the results of Example 3.1).

Answer Since $\Delta V = 100 \text{ V}$, $l = 1 \text{ m}$,

$$\therefore \text{electric field} = \frac{\Delta V}{l} = \frac{100}{1} = 100 \text{ Vm}^{-1}$$

Also, conductivity $\sigma = 5.81 \times 10^7 \Omega^{-1} \text{ m}^{-1}$

$$n = 8.5 \times 10^{28} \text{ m}^{-3}$$

$$\therefore v = \frac{\sigma E}{en} = \frac{5.81 \times 10^7 \times 100}{1.6 \times 10^{-19} \times 8.5 \times 10^{28}}$$

$$= 0.43 \text{ m s}^{-1}$$

$$v_{rms} = \sqrt{\frac{3k_B T}{m}}$$

$$= \sqrt{\frac{3 \times (1.38 \times 10^{-23} \text{ JK}^{-1}) \times 300 \text{ K}}{9.1 \times 10^{-31} \text{ kg}}}$$

$$= 1.17 \times 10^5 \text{ m s}^{-1}$$

3.5.1 Mobility

As we have seen, conductivity arises from mobile charge carriers. In metals, these mobile charge carriers are electrons; in an ionised gas, they

are electrons and positive charged ions; in an electrolyte, these can be both positive and negative ions. In a semiconductor material such as germanium or silicon, conduction is partly due to electrons and partly due to electron vacancies called holes. Holes are sites of missing electrons which act like positive charges (Chapter 15).

An important quantity is the mobility μ defined as the magnitude of the drift velocity per unit electric field

$$\mu = \frac{|v_d|}{E} \quad (3.20)$$

Mobility is positive for both — electrons and holes, although their drift velocities are opposite to each other. The electrical conductivity for a semiconductor containing electrons and holes as charge carriers can be expressed as:

$$\sigma = ne\mu_e + pe\mu_h \quad (3.21)$$

where μ_e and μ_h are electron and hole mobilities and n and p are electron and hole concentrations. Since

$$v_d = \frac{q\tau E}{m} \quad (3.22)$$

Also
$$\mu = \frac{v_d}{E} = \frac{q\tau}{m}$$

Therefore,

$$\mu_e = \frac{e\tau_e}{m_e} \quad (3.23)$$

and
$$\mu_h = \frac{e\tau_h}{m_h} \quad (3.24)$$

τ_e and τ_h are average relaxation time for electrons and holes, respectively. m_e and m_h refer to mass of electron and hole, respectively. Charge on either carrier is e .

Table 3.3

Mobilities of some materials, at room temperature, in cm^2/Vs

Materials	Electrons	Holes
Diamond	1800	1200
Silicon	1350	480
Germanium	3600	1800
InSb	800	450
GaAs	8000	300

SI unit of mobility is m^2/Vs and is 10^4 of the mobility in practical units (cm^2/Vs).

3.6 TEMPERATURE DEPENDENCE OF RESISTIVITY

The resistivity of all metallic conductors increases with temperature. Over a limited temperature range that is not too large, the resistivity of a metallic conductor can often be represented approximately by a linear relation:

$$\rho_T = \rho_0 [1 + \alpha(T - T_0)] \quad (3.25)$$

where ρ_0 is the resistivity at a reference temperature T_0 and ρ_T its value at temperature T . The factor α is called the *temperature coefficient of resistivity* and has dimensions of $(\text{temperature})^{-1}$. Typical values of ρ and α for some materials are given in Table 3.1. It may be noted that for elemental metals, α is $\sim 4 \times 10^{-3}$ per $^\circ\text{C}$. For these conductors, the temperature dependence of ρ at low temperatures is, however, non-linear as shown in Fig. 3.7(a). At low temperatures, the resistivity increases as a higher power of temperature [Fig. 3.7(a)].

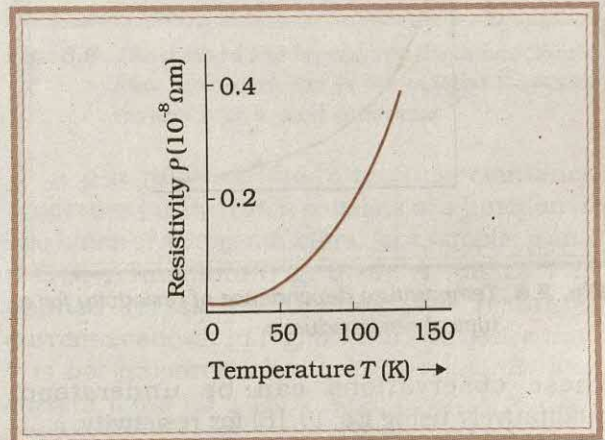


Fig. 3.7(a) Resistivity ρ_T of copper as a function of temperature T .

In a number of other metallic conductors, e.g., nichrome, which is an alloy of nickel, iron and chromium, the resistivity is very large but has a weak temperature dependence [Fig. 3.7(b)]. Similarly, the resistivity of alloy manganin is nearly independent of temperature. Nichrome has residual resistivity even at absolute zero, whereas a pure metal has vanishing (or very small) resistivity at absolute zero — a fact that can be used to check the purity of elemental metals.

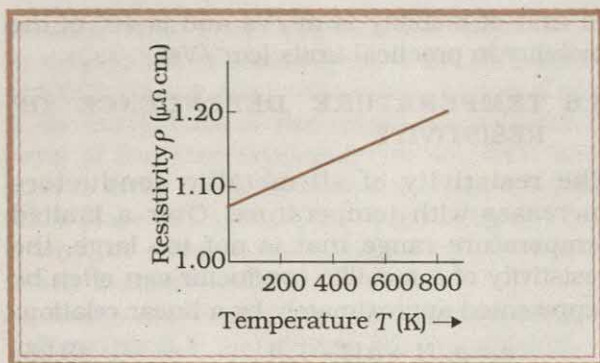


Fig. 3.7(b) Resistivity ρ_T of nichrome as a function of absolute temperature T .

The resistivity of carbon decreases with increasing temperature and its temperature co-efficient of resistivity is negative. The resistivity of a semiconductor decreases rapidly with increasing temperature [Fig. 3.8].

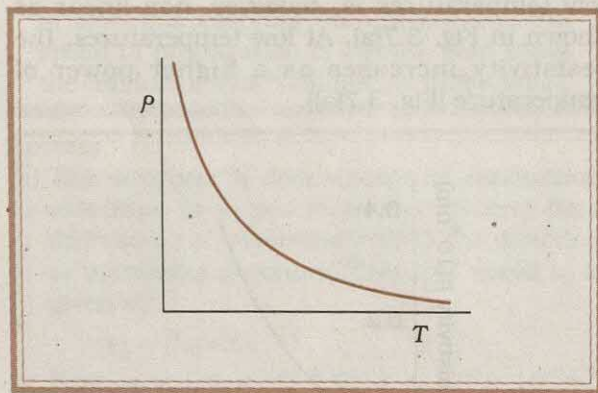


Fig. 3.8 Temperature dependence of resistivity for a typical semiconductor.

These observations can be understood qualitatively using Eq. (3.18) for resistivity:

$$\rho = \frac{m}{ne^2\tau}$$

1. For metals, the number of free electrons is fixed. As temperature increases, the amplitude of vibration of atoms/ions increases, and collisions of electrons with them become more effective and frequent, resulting in the decrease in τ and hence, increase in ρ . Thus for metals, ρ increases with temperature.
2. For insulators and semiconductors, it is not the relaxation time τ but the number density, n of free charge carriers that changes with

temperature. It can be shown that the number of charge carriers at temperature T is given by

$$n(T) = n_0 e^{E_g/k_B T} \quad (3.26)$$

where E_g is the energy gap that separates the valence and conduction bands in a solid (Chapter 15). At room temperature, $k_B T \approx 0.03$ eV. For metals, $E_g \approx 0$, so the number of charge carriers is fixed, nearly equal to $n_0 \approx 10^{28} \text{ m}^{-3}$. For semiconductors, E_g is about or somewhat less than 1 eV. For insulators, E_g may be considerably more than 1 eV. Correspondingly, resistivity of insulators at room temperature is very high, while semiconductors will have relatively less resistivity. Combining Eqs. (3.18) and (3.26), we get

$$\rho_T = \rho_0 e^{E_g/k_B T} \quad (3.27)$$

which shows that for semiconductors and insulators, resistivity increases with decreasing temperature.

Example 3.3 An electric toaster uses nichrome (an alloy of nickel, iron and chromium) for its heating element. When a negligibly small current passes through it, its resistance at room temperature (27.0°C) is found to be 75.3Ω . When the toaster is connected to a 230 V supply, the current settles after a few seconds to a steady value of 2.68 A. What is the steady temperature of the nichrome element? The temperature coefficient of resistance of nichrome averaged over the temperature range involved is $1.70 \times 10^{-4} \text{ }^\circ\text{C}^{-1}$.

Answer When the current through the element is very small, heating effects can be ignored and the temperature T_1 of the element is the same as room temperature. When the toaster is connected to the supply, its initial current will be slightly higher than its steady value of 2.68 A. But due to heating effect of the current, the temperature will rise. This will cause an increase in resistance and a slight decrease in current. In a few seconds, a steady state will be reached when temperature will rise no further, and both the resistance of the element and the current drawn will achieve steady values. The resistance R_2 at the steady temperature T_2 is

$$R_2 = \frac{230 \text{ V}}{2.68 \text{ A}} = 85.8 \Omega$$

Using the relation

$$R_2 = R_1 [1 + \alpha (T_2 - T_1)]$$

with $\alpha = 1.70 \times 10^{-4} \text{ } ^\circ\text{C}^{-1}$, we get

$$T_2 - T_1 = \frac{(85.8 - 75.3)}{(75.3) \times 1.70 \times 10^{-4}} = 820 \text{ } ^\circ\text{C}$$

$$\text{i.e., } T_2 = (820 + 27.0) \text{ } ^\circ\text{C} = 847 \text{ } ^\circ\text{C}$$

Thus, the steady temperature of the heating element (when heating effect due to the current equals heat loss to the surroundings) is $847 \text{ } ^\circ\text{C}$.

Example 3.4 The resistance of a platinum wire of platinum resistance thermometer at the ice point is 5Ω and at steam point is 5.39Ω . When the thermometer is inserted in a hot bath, the resistance of the platinum wire is 5.795Ω . Calculate the temperature of the bath?

Answer $R_0 = 5 \Omega$, $R_{100} = 5.23 \Omega$ and $R_t = 5.795 \Omega$

$$\begin{aligned} \text{Now, } t &= \frac{R_t - R_0}{R_{100} - R_0} \times 100, & R_t &= R_0 (1 + \alpha t) \\ &= \frac{5.795 - 5}{5.23 - 5} \times 100 \\ &= \frac{0.795}{0.23} \times 100 = 345.65 \text{ } ^\circ\text{C} \end{aligned}$$

3.7 LIMITATIONS OF OHM'S LAW

Ohm's law is not a fundamental law of nature, but a commonly found property of many substances under certain conditions. In many cases, therefore, the relation between voltage and current is different from that of Eq. (3.5). Some of these are specially worth mentioning.

There are a number of commonly used circuit elements with one or more of the following properties:

- V depends on I non-linearly.
- The relation between V and I depends on the sign of V for the same absolute value of V .
- The relation between V and I is non-unique, i.e., for the same current I , there is more than one value of voltage V .

We illustrate these with examples now. If the current flowing through a conductor is increased, the conductor becomes hotter and its resistance increases. But even if the temperature is kept constant, some conductors show an increase in resistivity as current increases (Fig. 3.9). This is genuine non-linear behaviour, easiest to see in p-n junctions, briefly described below. A detailed discussion of their nature and characteristics is given in Chapter 15.

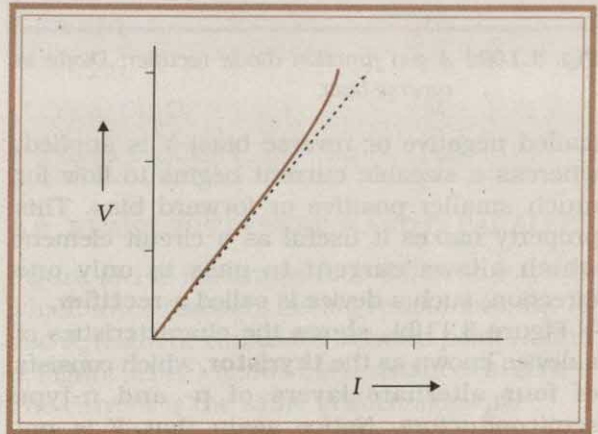


Fig. 3.9 The dashed line represents the linear Ohm's law. The solid line is the voltage V versus current I for a good conductor.

A p-n junction diode rectifier combines properties (a) and (b). It consists of a junction of two kinds of semiconductors, for example, p and n type germanium (Fig. 3.10). A voltage V is applied across the junction. The resulting current is shown in Fig. 3.11(a). We notice that V is not proportional to I . Further, hardly any current flows till a fairly large negative voltage

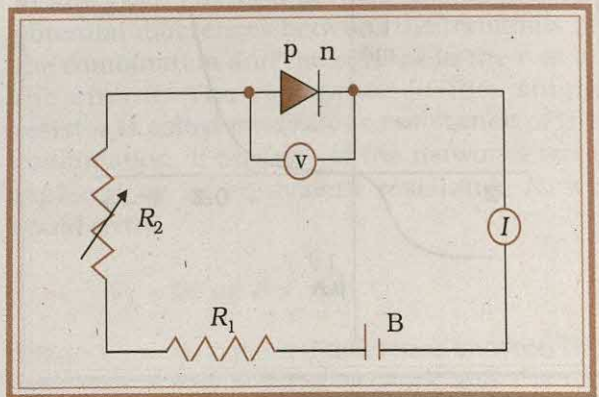


Fig. 3.10(a) A p-n junction diode rectifier: Diode in forward bias.

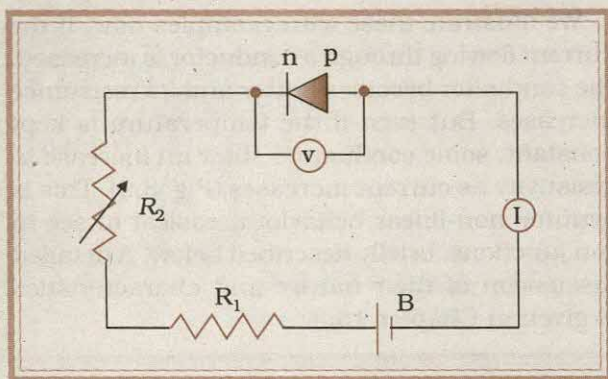


Fig. 3.10(b) A p-n junction diode rectifier: Diode in reverse bias.

(called negative or reverse bias) V is applied, whereas a sizeable current begins to flow for much smaller positive or forward bias. This property makes it useful as a circuit element which allows current to pass in only one direction; such a device is called a **rectifier**.

Figure 3.11(b) shows the characteristics of a device known as the **thyristor**, which consists of four alternate layers of p- and n-type semiconductors. Notice again that V is not proportional to I . All the features (a), (b) and (c) can be seen in the graph. The region AB is interesting because here the current carried by the device increases as the voltage decreases!

A more subtle and fundamental breakdown of Ohm's law, or more precisely the relations in Eqs. (3.7) and (3.18) occur in certain alloys at

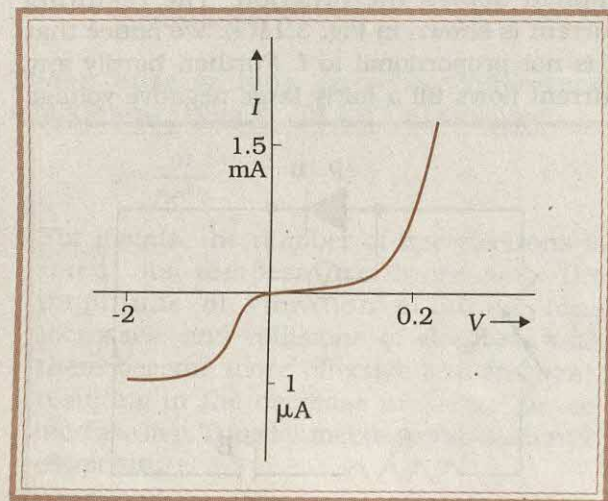


Fig. 3.11(a) Characteristic curve of a diode. Note the different scales for negative and positive values of the voltage and current.

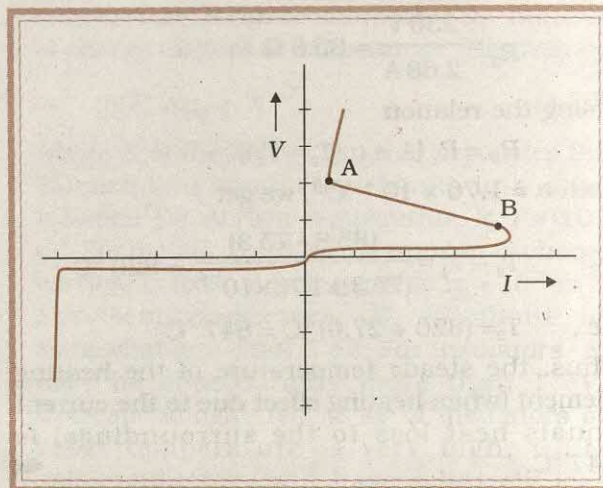


Fig. 3.11(b) Characteristics of a thyristor.

very low temperatures of order 4 K or less. It is found, for example, that the resistance of a wire of constant cross-section and the same material with length $2l$ is more than twice the resistance of the wire with length l . We have thought of electron flow as similar to flow of a viscous liquid in a tube under a pressure gradient (Poiseuille's experiment, Chapter 9, Class XI). However, electrons have a wave character (Chapter 12). This leads to interference effects, increasing the resistance with respect to the value it would have, if electrons flowed like water.

GaAs is a material in which non-linearity is more apparent. In fact, it is as much as after a certain voltage, the current decreases as voltage increases as shown in Fig. 3.12. This means that if ΔV is positive then ΔI is negative and hence effective resistance has the

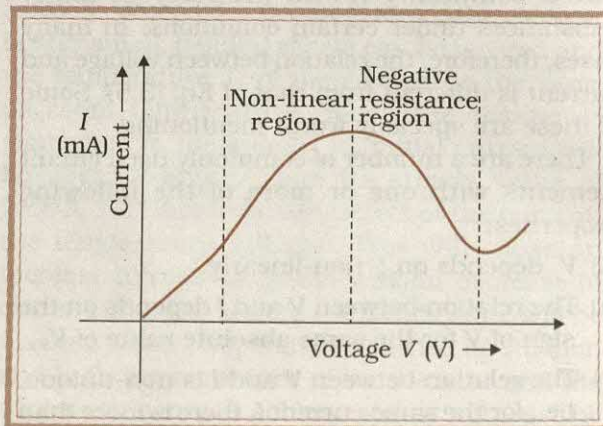


Fig. 3.12 Variation of current versus voltage for GaAs.

“conceptual meaning” of negative resistance. The reason for this lies in the special type of band gaps (both direct and indirect band gap) which is present in this material. You would learn about band gap in semiconductors in Chapter 15.

3.8 SUPERCONDUCTIVITY

The electrical resistivity of many metals and alloys drops suddenly to zero when their specimen are cooled to a sufficiently low temperature called the **critical temperature** (T_c). This phenomenon is called **superconductivity** and the material showing such a behaviour is called **superconductor**. The phenomenon was first discovered by H. Kamerlingh Onnes in 1911 in mercury and the critical temperature of Hg was found out to be 4.2 K. An electric current set up in a superconductor persists for a very long time. A superconductor shows Meissner effect discussed below:

3.8.1 Meissner Effect

Meissner and Ochsenfeld, in 1933, found that if a conductor is cooled in a magnetic field to below the transition temperature, then at the transition the lines of induction B are pushed out as shown in the Fig. 3.13.

The Meissner effect shows that a bulk superconductor behaves as if inside the specimen, B is zero.

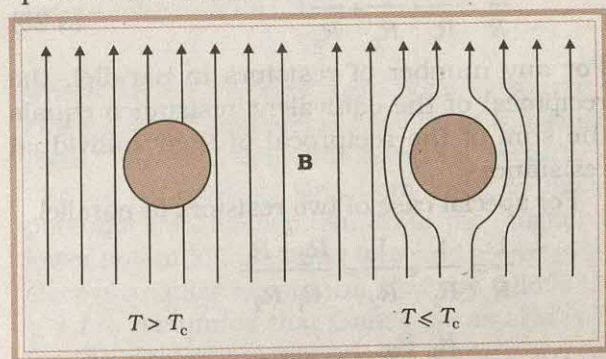


Fig. 3.13 Meissner effect in a superconductor.

3.8.2 High T_c Superconductivity

Superconductivity in some materials, mainly oxides, occurs upto much higher temperatures (Table 3.4). In 1988, 23 K critical temperature in inter metallic compounds was raised to 125 K in bulk superconducting oxides with properties conforming to Meissner effect, persistent currents and substantially zero dc resistivity. By 1994, these materials showed promise of

commercial applications in thin film devices, levitated vehicles and long distance power transmission. To date, these have not been very successful.

Table 3.4
Critical Temperature T_c (K) of some Superconducting Materials

Material	T_c (K)
Hg	4.2
Au_2Bi	1.7
Pb_2Au	7.0
$\text{YBa}_2\text{Cu}_3\text{O}_7$	90
$\text{Tl}_2\text{Ba}_2\text{Ca}_2\text{Cu}_3\text{O}_{10}$	120

3.9 RESISTORS IN SERIES AND IN PARALLEL

Figure (3.14) illustrates four different ways in which three resistors having resistances, R_1 , R_2 and R_3 might be connected between points x and y . Figure 3.14(a) shows the resistances in series. The current is the same in each element.

The resistors in Fig. 3.14(b) are said to be in parallel between the points x and y . Each resistor provides an alternative path between the points. The potential difference is the same across each element.

In Fig. 3.14(c) resistors, R_2 and R_3 are in parallel with each other and this combination is in series with resistance R_1 . In Fig. 3.14(d), R_2 and R_3 are in series and this combination is in parallel with R_1 .

It is always possible to find a single resistor that could replace a combination of resistors in any given circuit and leave unchanged the potential differences between the terminals of the combination and the current in the rest of the circuit. The resistance of this single resistor is called **equivalent resistance** of the combination. If any one of the networks were replaced by its equivalent resistance R , we could write

$$V_{xy} = IR \text{ or } R = \frac{V_{xy}}{I}$$

where V_{xy} is the potential difference between the terminals x and y of the network and I is the current at point x or y . Hence, the method of computing an equivalent resistance is to assume

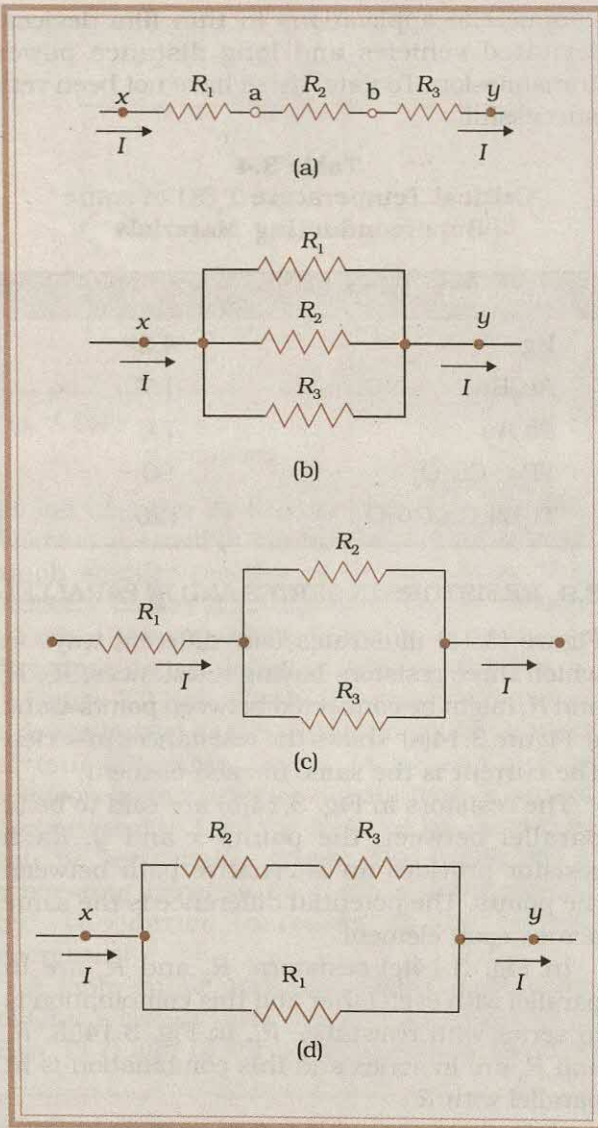


Fig. 3.14 Some combinations of resistors.

a potential difference V_{xy} across the actual network, compute the corresponding current I , and take the ratio V_{xy}/I .

If the resistors are in series as in Fig. 3.14(a), current in each one must be the same and equal to the line current I .

Hence,

$$V_{xa} = IR_1, V_{ab} = IR_2 \text{ and } V_{by} = IR_3$$

$$V_{xy} = V_{xa} + V_{ab} + V_{by}$$

$$= I(R_1 + R_2 + R_3)$$

$$\frac{V_{xy}}{I} = R_1 + R_2 + R_3$$

But $\frac{V_{xy}}{I}$ is, by definition, the equivalent resistance R . Therefore

$$R = R_1 + R_2 + R_3 \quad (3.28)$$

The equivalent resistance of any number of resistors in series equals the sum of their individual resistances.

If the resistors are in parallel as in Fig. 3.14(b), the potential difference between the terminals of each must be same and equal to V_{xy} . If the current in each are denoted by I_1 , I_2 and I_3 , respectively, then

$$I_1 = \frac{V_{xy}}{R_1}, I_2 = \frac{V_{xy}}{R_2}, I_3 = \frac{V_{xy}}{R_3}$$

Since charge is not accumulated at x , it follows that

$$I = I_1 + I_2 + I_3 = V_{xy} \left(\frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3} \right)$$

or
$$\frac{I}{V_{xy}} = \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3}$$

But $\frac{I}{V_{xy}} = \frac{1}{R}$, so that

$$\frac{1}{R} = \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3} \quad (3.29)$$

For any number of resistors in parallel, the reciprocal of the equivalent resistance equals the sum of the reciprocal of their individual resistances.

For special case of two resistors in parallel,

$$\frac{1}{R} = \frac{1}{R_1} + \frac{1}{R_2} = \frac{R_2 + R_1}{R_1 R_2}$$

and
$$R = \frac{R_1 R_2}{R_1 + R_2}$$

Also, since $V_{xy} = I_1 R_1 = I_2 R_2$

$$\frac{I_1}{I_2} = \frac{R_2}{R_1}$$

and the current carried by two resistors in parallel are inversely proportional to their resistances.

The equivalent resistance of the networks in Fig. 3.14(c) and Fig. 3.14(d) could be found by

the same general method, but it is simpler to consider them as combination of series and parallel arrangements. Thus, in Fig. 3.14(c) the combination of R_2 and R_3 in parallel is first replaced by its equivalent resistance which then forms a simple series combinations with R_1 . In Fig. 3.14(d), the combination of R_2 and R_3 in series forms a simple parallel combination with R_1 . Not all networks, however, can be reduced to simple series-parallel combinations and special methods must be used for handling such networks.

3.10 ELECTRIC CIRCUITS AND KIRCHHOFF'S RULES

Most electrical circuits consist not merely a single source and a single external resistor, but comprise of a number of sources, resistors or other elements such as capacitors, motors, etc. interconnected in a complicated manner. The general term applied to such a circuit is called *network*.

Here we start the discussion with internal resistance of electrical circuits and after going through the intricacies of resistance in series and parallel, we show how Kirchhoff's rules based on charge neutrality in a metal, greatly help in calculating electrical properties.

3.10.1 Internal Resistance

Let us return to Eq. (3.3)

$$\varepsilon = V + W'$$

where part of the total work done by source of emf per unit charge (ε) is used up in the source itself (W'). In a cell, this happens because in moving the positive ions from lower to higher potential (or negative ions from the higher to lower potential), the other ions and atoms of the electrolyte offer resistance. Now, by Ohm's law $V = IR$. Assuming that Ohm's law is also valid for the flow of current in the source, we can assign an internal resistance r to the source and write $W' = Ir$, so that Eq. (3.3) is written as

$$\varepsilon = V + Ir = IR + Ir = I(R + r)$$

This can also be written as

$$V = \varepsilon - Ir$$

showing that the external voltage is less than the emf of the source by the quantity Ir . It is as if an internal resistance r combines in series with the external resistance R to determine the current in the circuit for a given source of emf.

Clearly, if $I = 0$, i.e., the circuit is open, $V = V_{\text{open}}$ and we get Eq. (3.4).

3.10.2 Kirchhoff's Rules

We notice that the Eq. (3.29)

$$\frac{1}{R} = \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3} + \dots + \frac{1}{R_n}$$

uses the fact that there is no net current at the junction x , i.e., the incoming current equals the outgoing current [Fig. 3.14(b)]. Also, the potential difference across any resistor is the same, or if we complete the circuit $xy \rightarrow yx$ via a path involving any two resistors, the total potential change is zero. These facts, called Kirchhoff's rules, are very useful for many electrical circuit problems; they are discussed in detail below.

- (a) *First or junction rule:* This rule is based on the fact that charge cannot accumulate at any point in a conductor in a steady situation. It states that 'at any junction of several circuit elements, the sum of currents entering the junction must equal the sum of currents leaving it'. Unless this happens, net positive or negative charge will accumulate at the junction at a rate equal to the net electrical current at the junction. Consider for example, the circuit in Fig. 3.15. At the junction a , I_3 flows in and I_1 and I_2 flow out. We must then have

$$I_3 = I_1 + I_2 \quad (3.30)$$

Applying this to another junction of this circuit leads to nothing new. For further progress in analysing this circuit, we need another rule, given below.

- (b) *Second or loop rule:* 'The algebraic sum of changes in potential around any closed resistor loop must be zero'. Otherwise, one can continuously gain energy by circulating charge around a closed loop in a particular direction. So, this rule is based on energy conservation. Now consider the loop 'ahdcba' in Fig. 3.15.

We have, from Kirchhoff's second rules,

$$-30I_1 - 41I_3 + 45 = 0 \quad [3.31(a)]$$

For the second loop, the circuit 'ahdefga' is taken. We have

$$-30I_1 + 21I_2 - 80 = 0 \quad [3.31(b)]$$

These three relations [Eqs. (3.30) to (3.31)] can be used to calculate I_1 , I_2 and I_3 to be -0.86 A, 2.59 A and 1.73 A, respectively.

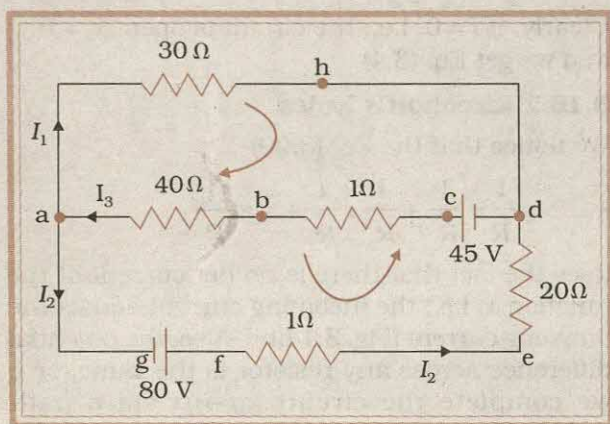


Fig. 3.15 An electric circuit.

Thus, the Kirchhoff's rules enable us to determine the currents and voltages in different parts of the electrical circuit.

3.10.3 Sign Convention in Applying Kirchhoff's Rules

The principal difficulty is not in understanding the basic ideas but in keeping track of algebraic signs. The following procedure should be followed carefully.

First, all quantities, known and unknown, should be labelled carefully, including an assumed sense of directions for each unknown current or emf. The solution is then carried out using the assumed directions, and if the actual direction of a particular quantity is opposite to the assumed direction, the value of the quantity will emerge from the analysis with a negative sign. Hence, Kirchhoff's rules, correctly used, give the direction and magnitude of unknown currents and emf's.

Usually in labelling currents it is advantageous to use the point rule immediately to express the currents in terms of as few quantities as possible. For example, Fig. 3.16 (a) shows a circuit correctly labelled, and Fig. 3.16(b) shows the same circuit, relabelled by applying the point rule to point a to eliminate I_3 .

The following guidelines will help with the problems of signs:

1. Choose any closed loop in the network, and designate a direction (clockwise or counter clockwise) to transverse the loop in applying the loop rule.
2. Go around the loop in the designated direction, adding emf's and potential differences. An emf

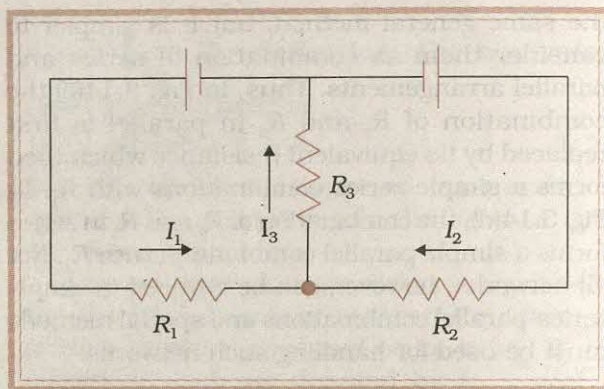


Fig. 3.16(a)

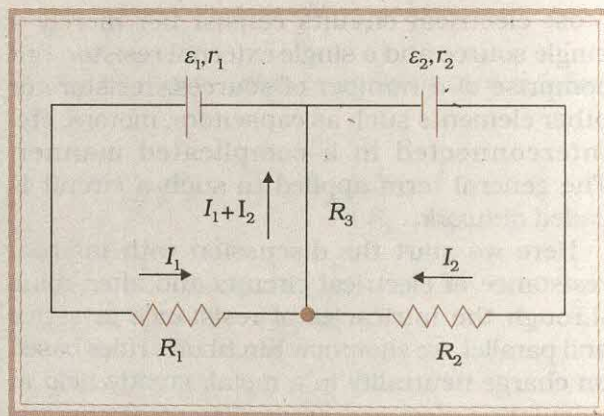


Fig. 3.16(b)

is counted as positive when it is traversed from (-) to (+) and negative when transformed from (+) to (-). An IR term is counted negative if the resistor is traversed in the same direction of the assumed current, and positive if in the opposite direction.

3. Equate the sum of step (2) to zero.
4. If necessary, choose another loop to obtain different relations between the unknowns, and continue until there are as many equations and unknowns or until every circuit element has been included in at least one of the chosen loops.

Example 3.5 A network of resistances is connected to a 16 V battery with internal resistance of 1Ω as shown in Fig. 3.17. (a) Compute the equivalent resistance of the network. (b) Obtain the current in each resistor and (c) Obtain the voltage drops V_{AB} , V_{BC} and V_{CD} .

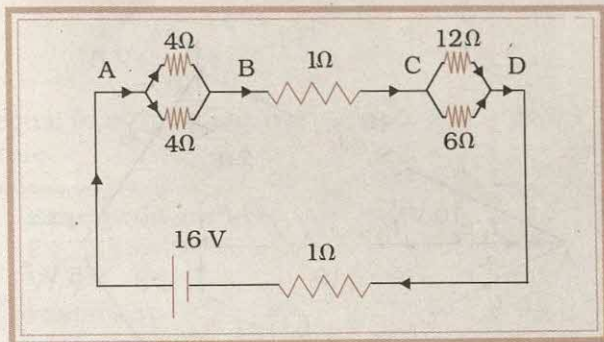


Fig. 3.17

Answer

- (a) The network is a simple series and parallel combination of resistances. First the two 4Ω resistors in parallel are equivalent to a resistance $= [(4 \times 4)/(4 + 4)]\Omega = 2\Omega$. In the same way, the 12Ω and 6Ω resistances in parallel are equivalent to a resistance of $[(12 \times 6)/(12 + 6)]\Omega = 4\Omega$.

The equivalent resistance R of the network is obtained by combining these resistances (2Ω and 4Ω) with 1Ω in series, i.e.,

$$R = 2\Omega + 4\Omega + 1\Omega = 7\Omega.$$

- (b) The total current I in the circuit is

$$I = \frac{\varepsilon}{R+r} = \frac{16V}{(7+1)\Omega} = 2A$$

Consider the resistors between A and B. If I_1 is the current in one of the 4Ω resistors and I_2 the current in the other.

$$I_1 \times 4 = I_2 \times 4$$

i.e., $I_1 = I_2$, which is otherwise obvious from the symmetry of the two arms. But $I_1 + I_2 = I = 2A$. Thus,

$$I_1 = I_2 = 1A$$

i.e., current in each 4Ω resistor is $1A$. Current in 1Ω resistor between B and C would be $2A$.

Now, consider the resistances between C and D. If I_3 is the current in the 12Ω resistor, and I_4 in the 6Ω resistor,

$$I_3 \times 12 = I_4 \times 6, \text{ i.e., } I_4 = 2I_3$$

$$\text{But } I_3 + I_4 = I = 2A$$

$$\text{Thus } I_3 = \left(\frac{2}{3}\right)A, I_4 = \left(\frac{4}{3}\right)A$$

i.e. the current in the 12Ω resistor is $(2/3)A$, while the current in the 6Ω resistor is $(4/3)A$.

- (c) The voltage drop across AB is

$$V_{AB} = I_1 \times 4 = 1A \times 4\Omega = 4V$$

which is also obtained by multiplying the total current between A and B by the equivalent resistance between A and B, i.e.,

$$V_{AB} = 2A \times 2\Omega = 4V$$

The voltage drop across BC is

$$V_{BC} = 2A \times 1\Omega = 2V$$

Finally, the voltage drop across CD is

$$V_{CD} = 12\Omega \times I_3 = 12\Omega \times \left(\frac{2}{3}\right)A = 8V$$

which is also obtained by multiplying total current between C and D by the equivalent resistance between C and D, i.e.,

$$V_{CD} = 2A \times 4\Omega = 8V$$

Note that the total voltage drop across AD is $4V + 2V + 8V = 14V$. Thus, the terminal voltage of the battery is $14V$, while its emf is $16V$. The loss of the voltage ($= 2V$) is accounted for by the internal resistance 1Ω of the battery $2A \times 1\Omega = 2V$.

Example 3.6 A battery of $10V$ and negligible internal resistance is connected across the diagonally opposite corners of a cubical network consisting of 12 resistors each of resistance 1Ω (Fig. 3.18). Determine the equivalent resistance of the network and the current along each edge of the cube.

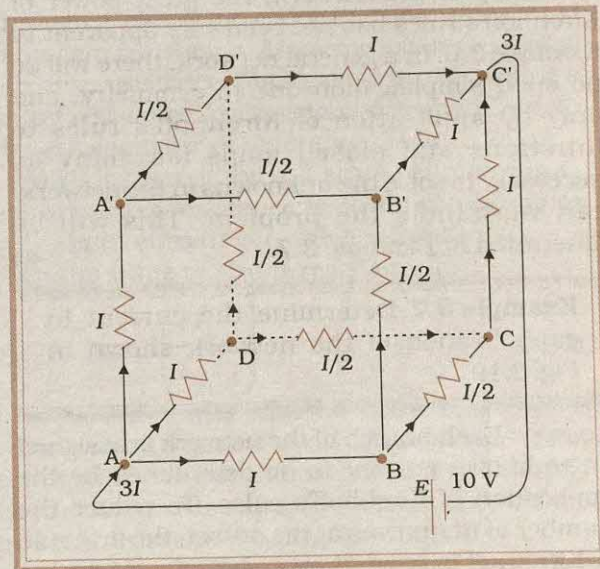


Fig. 3.18

Answer The network is not reducible to a simple series and parallel combinations of resistances. There is, however, a clear symmetry in the problem which we can exploit to obtain the equivalent resistance of the network.

The paths AA', AD and AB are obviously symmetrically placed in the network. Thus, the current in each must be the same, say, I . Further, at the corners A', B and D, the incoming current I must split equally into the two outgoing branches. In this manner, the current in all the 12 edges of the cube are easily written down in terms of I , using Kirchhoff's first rule and the symmetry in the problem.

Next take a closed loop, say, ABCC'EA, and apply Kirchhoff's second rule:

$$-IR - (1/2)IR - IR + \varepsilon = 0$$

where R is the resistance of each edge and ε the emf of battery. Thus,

$$\varepsilon = \frac{5}{2}IR$$

The equivalent resistance of the network R_{eq} is:

$$R_{eq} = \frac{\varepsilon}{3I} = \frac{5}{6}R$$

For $R = 1 \Omega$, $R_{eq} = (5/6) \Omega$ and for $\varepsilon = 10 \text{ V}$, the total current ($= 3I$) in the network is

$$3I = 10 \text{ V} / (5/6) \Omega = 12 \text{ A i.e., } I = 4 \text{ A}$$

The current flowing in each edge can now be read off from the Fig. 3.18.

It should be noted that because of the symmetry of the network, the great power of Kirchhoff's rules has not been very apparent in Example 3.6. In a general network, there will be no such simplification due to symmetry, and only by application of Kirchhoff's rules to junctions and closed loops (as many as necessary to solve the unknowns in the network) can we handle the problem. This will be illustrated in Example 3.7.

Example 3.7 Determine the current in each branch of the network shown in Fig. 3.19.

Answer Each branch of the network is assigned an unknown current to be determined by the application of Kirchhoff's rules. To reduce the number of unknowns at the outset, the first rule of Kirchhoff is used at every junction to assign the unknown current in each branch. We then

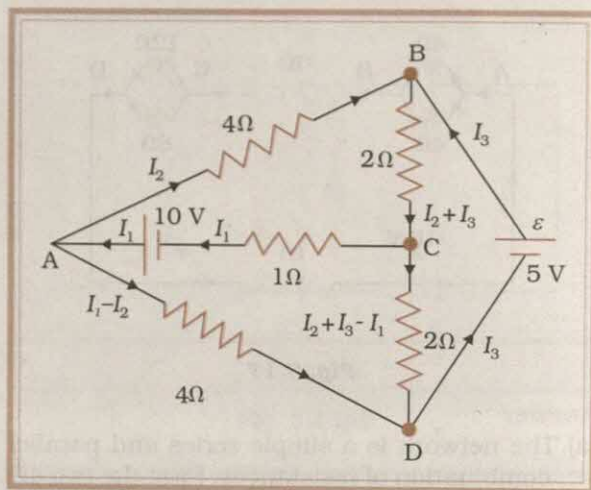


Fig. 3.19

have three unknowns I_1 , I_2 and I_3 which can be found by applying the second rule of Kirchhoff to three different closed loops. Kirchhoff's second rule for the closed loop ADCA gives:

$$10 - 4(I_1 - I_2) + 2(I_2 + I_3 - I_1) - I_1 = 0 \quad (a)$$

$$\text{i.e., } 7I_1 - 6I_2 - 2I_3 = 10$$

For the closed loop ABCA, we get

$$10 - 4I_2 - 2(I_2 + I_3) - I_1 = 0$$

$$\text{i.e., } I_1 + 6I_2 + 2I_3 = 10 \quad (b)$$

For the closed loop BCDEB, we get

$$5 - 2(I_2 + I_3) - 2(I_2 + I_3 - I_1) = 0$$

$$\text{i.e., } 2I_1 - 4I_2 - 4I_3 = -5 \quad (c)$$

Equations (a), (b) and (c) are three simultaneous equations in three unknowns. These can be solved by the usual method to give

$$I_1 = 2.5 \text{ A}, \quad I_2 = \frac{5}{8} \text{ A}, \quad I_3 = 1\frac{7}{8} \text{ A}.$$

The currents in the various branches of the network are:

$$AB : \frac{5}{8} \text{ A}, \quad CA : 2\frac{1}{2} \text{ A}, \quad DEB : 1\frac{7}{8} \text{ A},$$

$$AD : 1\frac{7}{8} \text{ A}, \quad CD : 0 \text{ A}, \quad BC : 2\frac{1}{2} \text{ A}.$$

It is easily verified that Kirchhoff's second rule applied to the remaining closed loops does not provide any additional independent equation; that is, the above values of currents satisfy the second rule for every closed loop of the network. For example, the total voltage drop over the closed loop BADEB is

$$5\text{ V} + \left(\frac{5}{8} \times 4\right)\text{ V} - \left(\frac{15}{8} \times 4\right)\text{ V}$$

equal to zero, as required by Kirchhoff's second rule.

Example 3.8 In the circuit shown in the Fig. 3.20(a), E, F, G and H are cells of emf 2 V, 1V, 3V and 1V, and their internal resistances are 2Ω , 1Ω , 3Ω and 1Ω , respectively. Calculate (a) the potential difference between B and D, (b) The potential difference across the terminals of each cells G and H.

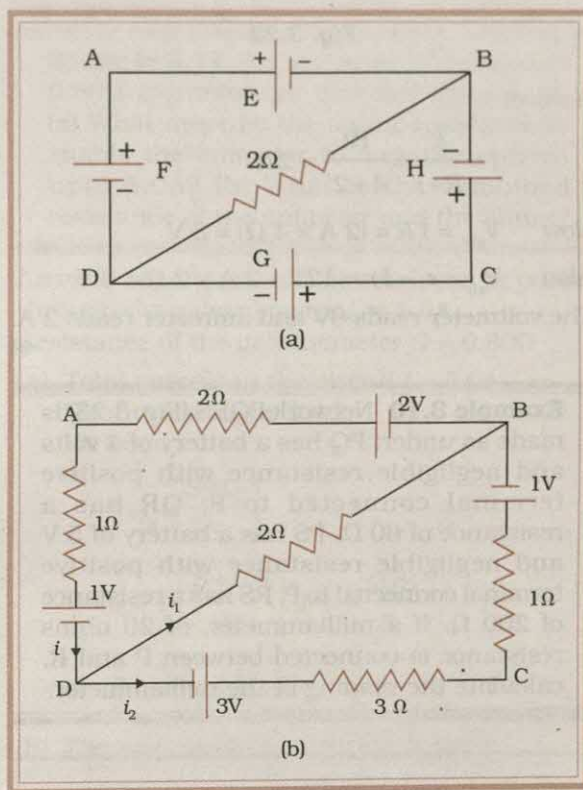


Fig. 3.20

Answer Figure 3.20(b) shows the network with emf and internal resistances explicitly.

(a) $i = i_1 + i_2$ (i)

Applying Kirchhoff's law to mesh ADBA

$$2 \times i + 1 \times i + 2 \times i_1 = 2 - 1 = 1$$

$$3i + 2i_1 = 1 \quad \text{(ii)}$$

Applying Kirchhoff's second law to mesh DCBD

$$3 \times i_2 + 1 \times i_2 - 2 \times i_1 = 3 - i$$

$$4i_2 - 2i_1 = 2 \quad \text{(iii)}$$

Solving (i), (ii) and (iii) we have,

$$i_1 = -\frac{1}{13}\text{ A}, \quad i_2 = \frac{6}{13}\text{ A}, \quad \text{and } i = \frac{5}{13}\text{ A}.$$

Potential difference between B and D

$$V_1 = i_1 \times 2 = \frac{2}{13}\text{ V}$$

(b) Potential difference across the terminals of cells G and H respectively

$$V_2 = \varepsilon - i_2 R = 3 - \frac{6}{13} \times 3 = 1.615\text{ V}$$

$$V_3 = \varepsilon - i_2 \times R = 1 - \frac{6}{13} \times 1 = 0.54\text{ V}$$

3.11 MEASUREMENT OF VOLTAGES, CURRENTS AND RESISTANCES

3.11.1 Voltmeter and Ammeter

Nothing has been mentioned for measuring the parameters such as current, voltage and resistance as yet. One basic instrument which helps in measuring the quantities is the galvanometer which produces a deflection proportional to electric current flowing through it. Using a galvanometer we can prepare instruments for measuring voltage and current called voltmeter and ammeter, respectively.

An ammeter is connected in series with the circuit as it is used to measure the current flowing through it. Inserting an ammeter in series with a circuit should not change the current flowing through it, clearly indicating that ammeter should have zero resistance. Ammeter has a very small effective resistance. This is done by connecting a shunt resistance (low resistance) S in parallel with the galvanometer of resistance G [Fig. 3.21(a)].

Now, if current for full scale deflection is I_g , galvanometer resistance is G , shunt resistance is S , shunt current is I_s , total current $I = (I_g + I_s)$, then the value of shunt resistance is

$$S = G \left(\frac{I_g}{I - I_g} \right) \quad (3.32)$$

A galvanometer can be converted into a voltmeter by connecting a high resistance R in series with the galvanometer coil [Fig. 3.21(b)]. High resistance is required to minimise the drop in total current flowing through the other circuits connected across it. For a high resistance connected and for a given voltage, the current

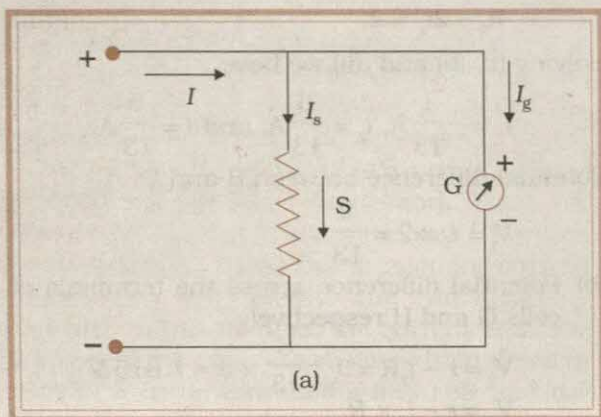


Fig. 3.21(a)

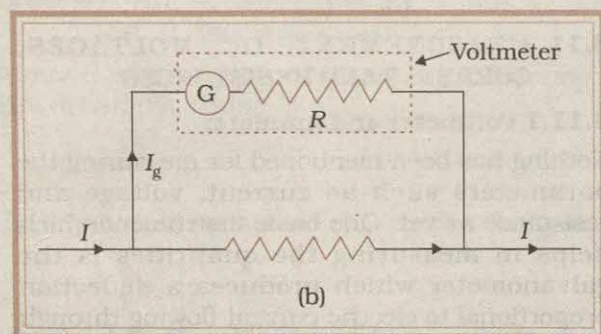


Fig. 3.21(b)

through the voltmeter is small.

$$I_g = V/(R + G)$$

$$R + G = V/I_g$$

Where I_g is the full scale deflection current, R is the high resistance in series with the galvanometer. I_g is the current flowing through the galvanometer

$$\text{or } R = \left(\frac{V}{I_g} - G \right) \quad (3.33)$$

Equations (3.32) and (3.33) form the basis for conversion of a galvanometer into an ammeter and a voltmeter respectively.

Example 3.9 In Fig. 3.22, an ammeter A and a resistor of resistance $R = 4 \Omega$ have been connected to the terminals of the source to form a complete circuit. The emf of the same is $12V$ having an internal resistance of 2Ω . Calculate the voltmeter and the ammeter reading.

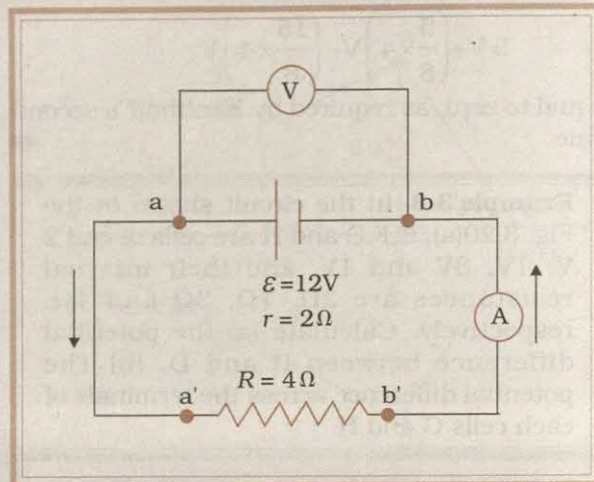


Fig. 3.22

Answer

$$I = \frac{\varepsilon}{R + r} = \frac{12}{4 + 2} = 2A$$

$$\text{Now } V_{a'b'} = IR = (2A \times 4\Omega) = 8V$$

$$\text{Also } V_{ab} = \varepsilon - Ir = 12V - 2A \times 2\Omega = 8V$$

The voltmeter reads $8V$ and ammeter reads $2A$.

Example 3.10 Network PQRS (Fig. 3.23) is made as under: PQ has a battery of 4 volts and negligible resistance with positive terminal connected to P ; QR has a resistance of 60Ω . PS has a battery of 5 V and negligible resistance with positive terminal connected to P , RS has a resistance of 200Ω . If a millimeter, of 20Ω resistance is connected between P and R , calculate the reading of the millimeter.

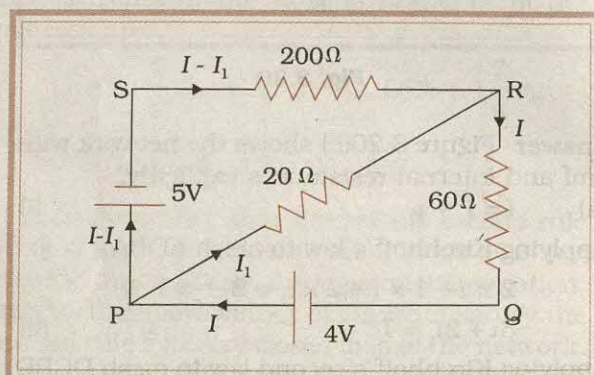


Fig. 3.23

Answer Applying Kirchhoff's law to PQRP loop

$$4 = 20 I_1 + 60 I \quad (i)$$

Considering PSRP loop

$$-5 = 200 (I - I_1) - 20 I_1,$$

$$-1 = 40 I - 40 I_1 - 4 I_1,$$

$$40 I - 44 I_1 = -1 \quad (ii)$$

Multiplying (ii) by 3, we have

$$120 I - 132 I_1 = -3 \quad (iii)$$

Multiplying (i) by 2, we have

$$120 I + 40 I_1 = 8 \quad (iv)$$

Solving (iii) and (iv), we get

$$I_1 = 11/172 = 0.064 \text{ A}$$

Example 3.11 An ammeter of resistance 0.80Ω can measure currents upto 1.0A .
(a) What must be the shunt resistance to enable the ammeter to measure current upto 5.0A ? (b) What is the combined resistance of the ammeter and the shunt?

Answer Maximum current which can be passed through the galvanometer $I_g = 1.0\text{A}$.
Resistance of the galvanometer $G = 0.80\Omega$

(a) Total current in the circuit $I = 5.0\text{A}$

Shunt resistance $S = ?$

We know

$$I = I_g \left(\frac{S+G}{S} \right)$$

$$5 = 1 \left(\frac{S+0.8}{S} \right) = 1 + \frac{0.8}{S}$$

$$\text{or } S = \frac{0.8}{4} = 0.2 \Omega$$

(b) The combined resistance will be

$$\frac{1}{R} = \frac{1}{0.8} + \frac{1}{0.2}$$

$$\frac{1}{R} = \frac{1+4}{0.8}$$

$$\text{or } R = 0.16 \Omega.$$

3.11.2 Wheatstone Bridge

Figure 3.24 shows the fundamental diagram of wheatstone bridge. The bridge has four resistive arms, together with a source of emf (a battery) and a galvanometer which is a null detector. The

current through the galvanometer depends on the potential difference between the point c and d. The bridge is said to be balanced when the potential difference across the galvanometer is 0V so that there is no current through the galvanometer. This condition occurs when the voltage from point c to point a, equals the voltage from point d to point a; or by referring to the other battery terminal, when the voltage from point c to point b equals the voltage from point d to point b. Hence, the bridge is balanced when

$$I_1 R_1 = I_2 R_2 \quad (3.34)$$

If the galvanometer current is zero, the following conditions also exist:

$$I_1 = I_3 = \frac{\epsilon}{R_1 + R_3} \quad (3.35)$$

$$\text{and } I_2 = I_4 = \frac{\epsilon}{R_2 + R_4} \quad (3.36)$$

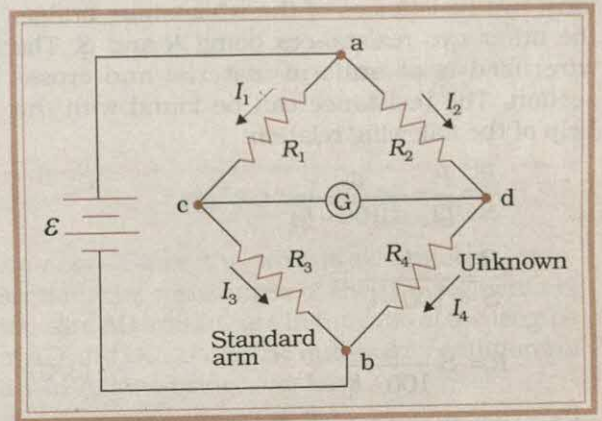


Fig. 3.24 Wheatstone Bridge.

Combining Eqs. (3.34), (3.35) and (3.36) and simplifying, we obtain

$$\frac{R_1}{R_1 + R_3} = \frac{R_2}{R_2 + R_4} \quad (3.37)$$

from which we get

$$R_1 R_4 = R_2 R_3 \quad (3.38)$$

Equation (3.38) is the well known expression for balance of the wheatstone bridge. If three of the resistances have known value, the fourth may be determined from Eq. (3.38). Hence, if R_4 is the unknown resistor, its resistance can be expressed in terms of remaining resistors

$$R_4 = R_3 \frac{R_2}{R_1} \quad (3.39)$$

Resistance R_3 is called the standard arm of the bridge and resistors R_2 and R_1 are called the ratio arms.

3.11.3 Metre Bridge — Special Case of Wheatstone Bridge

This is the simplest form of wheatstone bridge and is specially useful for comparing resistances more accurately. The construction of the metre bridge is shown in the Fig. 3.25. It consists of one metre resistance wire clamped between two metallic strips bent at right angles and it has two points for connection. There are two gaps; in one of them a known resistance whose value is to be determined is connected. The galvanometer is connected with the help of jockey across BD and the cell is connected across AC. After making connections, the jockey is moved along the wire and the null point is obtained. The segment of length l_1 and $(100 - l_1)$ form two resistances of the wheatstone bridge, the other two resistances being R and S . The wire used is of uniform material and cross-section. The resistance can be found with the help of the following relation:

$$\frac{R}{S} = \frac{P}{Q} = \frac{\sigma l_1}{\sigma(100 - l_1)}$$

$$\frac{R}{S} = \frac{l_1}{100 - l_1}$$

$$R = S \frac{l_1}{100 - l_1} \quad (3.40)$$

where σ is the resistance per unit length of the wire and l_1 is the length of the wire from one end where null point is obtained. The bridge is most sensitive when null point is somewhere near the middle point of the wire. This is due to end resistances.

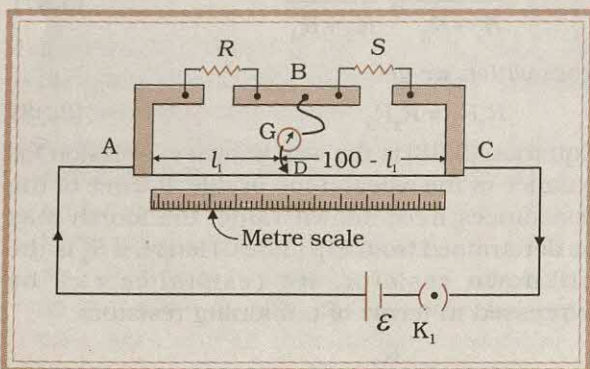


Fig. 3.25

End correction: Sometimes at the end points of the wire, some length is found under the metallic strips and as a result, in addition to length l_1 or $(100 - l_1)$ some additional length should be added for accurate measurements. The resistance due to this additional length is called *end resistance*. If the end resistance is small it can be determined by first introducing known resistances P and Q in the gap and obtaining the null point reading l_1 , then interchanging P and Q and obtaining the null point reading l_2 . Let α and β be the lengths on the respective end under the metallic strips, then we have

$$\frac{P}{Q} = \frac{l_1 + \alpha}{100 - l_1 + \beta} \quad (3.41)$$

$$\frac{Q}{P} = \frac{l_2 + \alpha}{100 - l_2 + \beta} \quad (3.42)$$

Solving Eqs. (3.41) and (3.42) for α and β , we have

$$\alpha = \frac{Ql_1 - Pl_2}{P - Q} \quad (3.43)$$

$$\beta = \frac{Pl_1 - Ql_2}{P - Q} - 100 \quad (3.44)$$

Hence the value of α , β can be calculated and suitably accounted for when accurate measurements are required.

Example 3.12 The four arms of a wheatstone bridge (Fig. 3.26) have the following resistances:

$AB = 100 \Omega$, $BC = 10 \Omega$, $CD = 5 \Omega$ and $DA = 60 \Omega$.

The galvanometer of 15Ω resistance is connected across BD. Calculate the current through the galvanometer when at potential difference of 10 V is maintained across AC.

Answer Considering the mesh BADB, we have

$$100I_1 + 15I_g - 60I_2 = 0$$

$$\text{or } 20I_1 + 3I_g - 12I_2 = 0 \quad (i)$$

Considering the mesh BCDB, we have

$$10(I - I_g) - 15I_g - 5(I_2 + I_g) = 0$$

$$10I_1 - 30I_g - 5I_2 = 0$$

$$2I_1 - 6I_g - I_2 = 0 \quad (ii)$$

Considering the mesh ADCEA,

$$60I_2 + 5(I_2 + I_g) = 10$$

$$65I_2 + 5I_g = 10$$

$$13I_2 + I_g = 2$$

(iii)

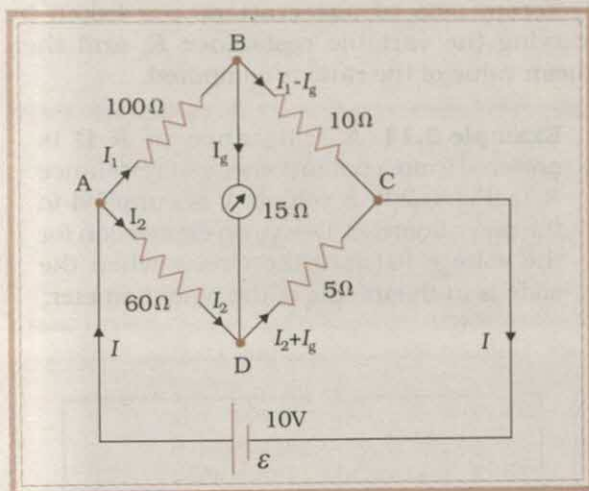


Fig. 3.26

Multiplying Eq. (iii) by 10

$$20I_1 - 60I_g - 10I_2 = 0$$

From (iv) and (i) we have

$$63I_g - 2I_2 = 0$$

$$I_2 = \frac{63}{2} I_g = 31.5I_g \quad (v)$$

Substituting the value of I_2 into Eq. (iii), we get

$$13(31.5I_g) + I_g = 2$$

$$410.5 I_g = 2$$

$$I_g = 4.87 \text{ mA.}$$

Example 3.13 Figure 3.27(a) shows a metre bridge (which is nothing but a practical Wheatstone bridge) consisting of two resistors X and Y together in parallel with a metre long constantan wire of uniform cross section. With the help of a movable contact D , one can change the ratio of the resistances of the two segments of the wire, until a sensitive galvanometer G connected across B and D shows no deflection. The null point is found to be at a distance of 33.7 cm from the end A . The resistance Y is shunted by a resistance Y' of 12.0Ω [Fig. 3.27(b)] and the null point is found to shift by a distance of 18.2 cm. Determine the resistance of X and Y .

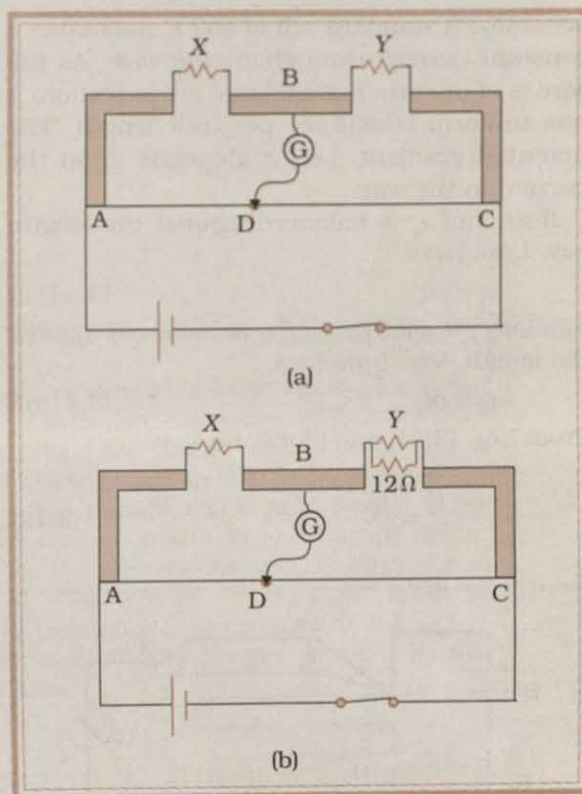


Fig. 3.27(a) and (b)

Answer Since the wire is of uniform cross-section, the resistances of the two segments of the wire AD and DC are in the ratio of the lengths of AD and DC . Using the null-point conditions of a wheatstone bridge, we have

$$(X/Y) = (33.7/66.3) \quad (i)$$

When Y is shunted by a resistance of 12.0Ω , its net resistance changes

$$Y' = 12Y(Y + 12)$$

Since Y' is less than Y , the ratio X/Y' is greater than X/Y . Thus, the null point must shift towards the end C , i.e.,

$$(X/Y') = (51.9/48.1)$$

$$\text{or } X(Y + 12)/12Y = (51.9/48.1)$$

$$\text{i.e., } \frac{Y + 12}{12} = \left(\frac{51.9}{48.1} \right) \times \frac{66.3}{33.7}$$

which gives $Y = 13.5 \Omega$ and $X = 6.86 \Omega$, using (i)

3.11.4 Potentiometer

In this instrument the resistance wire is of more than a metre in length. This enables greater

accuracy. A standard cell of emf ε_1 maintains a constant current throughout the wire. As the wire is of uniform material and cross-section, it has uniform resistance per unit length. The potential gradient, i.e., ρ , depends upon the current in the wire.

If an emf ε_1 is balanced against the length, say, l_1 we have

$$\varepsilon_1 = \rho l_1 \quad [3.41(a)]$$

Similarly, if another emf ε_2 is balanced against the length, say, l_2 we have

$$\varepsilon_2 = \rho l_2 \quad [3.41(b)]$$

From Eqs. (3.41) and (3.42), we have

$$\frac{\varepsilon_1}{\varepsilon_2} = \frac{l_1}{l_2} \quad (3.42)$$

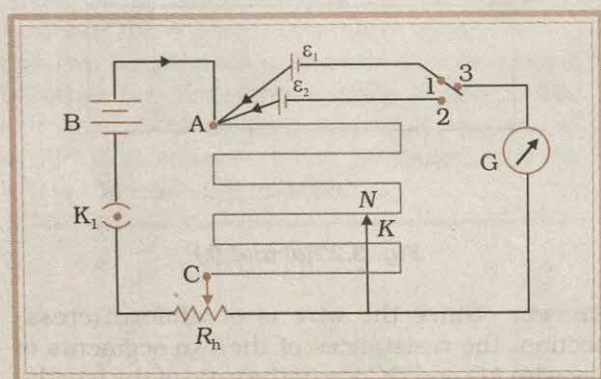


Fig. 3.28

From Fig. 3.28, by means of a battery B and rheostat R_h a steady current is passed through the potentiometer wire AC. Two cells ε_1 and ε_2 whose emf's are to be compared are put in such a way that positive terminals are connected to A and negative terminal to the galvanometer through a two-way plug key.

First the cell ε_1 is connected by connecting 1 and 3 points of key and by moving the jockey K on the potentiometer wire, the no deflection point is obtained. Let the reading be l_1 , then

$$\varepsilon_1 = \rho l_1$$

where ρ is the potential gradient and l_1 is the length AN. After this, the points 2 and 3 of the key are connected i.e., the cell of emf ε_2 is put into the circuit and again the no deflection point on the wire is obtained. Let this reading be l_2 . Then

$$\varepsilon_2 = \rho l_2$$

$$\frac{\varepsilon_1}{\varepsilon_2} = \frac{l_1}{l_2}$$

Different sets of observations are taken by varying the variable resistance R_k and then mean value of the ratio is computed.

Example 3.14 A resistance of $R \Omega$ is powered from a potentiometer of resistance $R_0 \Omega$ (Fig. 3.29). A voltage V is supplied to the potentiometer. Derive an expression for the voltage fed into the circuit when the slide is in the middle of the potentiometer.

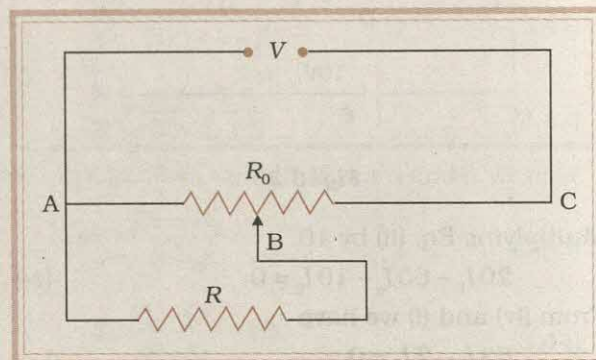


Fig. 3.29

Answer While the slide is in the middle of the potentiometer only half of its resistance ($R_0/2$) will be between the points A and B. Hence, the total resistance between A and B, say, R_1 , will be given by the following expression:

$$\frac{1}{R_1} = \frac{1}{R} + \frac{1}{(R_0/2)}$$

$$R_1 = \frac{R R_0}{R_0 + 2R}$$

The total resistance between A and C will be sum of resistance between A and B and B and C, i.e., $R_1 + R_0/2$

\therefore The current flowing through the potentiometer will be

$$I = \frac{V}{R_1 + R_0/2} = \frac{2V}{2R_1 + R_0}$$

The voltage V_1 taken from the potentiometer will be the product of current I and resistance R_1 ,

$$V_1 = I R_1 = \left(\frac{2V}{2R_1 + R_0} \right) \times R_1$$

Substituting for R_1 , we have

$$V_1 = \frac{2V}{2\left(\frac{R_s \times R}{R_s + 2R}\right) + R_s} \times \frac{R_s \times R}{R_s + 2R}$$

$$V_1 = \frac{2VR}{2R + R_s + 2R}$$

$$\text{or } V_1 = \frac{2VR}{R_s + 4R}$$

SUMMARY

1. Current through a given area of a conductor is the net charge passing per unit time through the area.
2. To maintain a steady current, we must have a closed circuit in which an external agency moves electric charge from lower to higher potential energy. The work done per unit charge by the source in taking the charge from lower to higher potential energy (i.e., from one terminal of the source to the other) is called the electromotive force, or *emf*, of the source. Note that the emf is not a force; it is the voltage difference between the two terminals of a source in open circuit.
3. *Ohm's law*: The electric current I flowing through a substance is proportional to the voltage V across its ends, i.e., $V \propto I$ or $V = RI$, where R is called the resistance of the substance. The unit of resistance is ohm: $1\Omega = 1 \text{ V A}^{-1}$.
4. The resistance R of a conductor depends on its length l and constant cross-sectional area A through the relation,

$$R = \frac{\rho l}{A}$$

where ρ , called resistivity is a property of the material and depends on temperature and pressure.

5. Electrical resistivity of substances varies over a very wide range. Metals have low resistivity, in the range of $10^{-8} \Omega \text{ m}$ to $10^{-6} \Omega \text{ m}$. Insulators like glass and rubber have 10^{22} to 10^{24} times greater resistivity. Semiconductors like Si and Ge lie roughly in the middle range of resistivity on a logarithmic scale.
6. In most substances, the carriers of current are electrons; in some cases e.g., ionic crystals and electrolytic liquids, positive and negative ions carry the electric current.
7. Current density \mathbf{j} gives the amount of charge flowing per second per unit area normal to the flow.

$$\mathbf{j} = nq v_d$$

where n is the number density (number per unit volume) of charge carriers each of charge q , and v_d is the drift velocity of the charge carriers. For electrons $q = -e$. If \mathbf{j} is normal to a cross-sectional area \mathbf{A} and is constant over the area, the magnitude of the current I through the area is $(nev_d A)$.

8. Using $E = V/l$, $I = nev_d A$, and Ohm's law, one obtains

$$\frac{eE}{m} = \rho \frac{ne^2}{m} v_d$$

The proportionality between the force eE on the electrons in a metal due to the external field E and the drift velocity v_d (not acceleration) can be understood, if we assume that the electrons suffer collisions with ions in the metal, which deflect them randomly. If such collisions occur on an average at a time interval τ ,

$$v_d = a\tau = eE\tau/m$$

where a is the acceleration of the electron. This gives

$$\rho = \frac{m}{ne^2\tau}$$

9. In the temperature range in which resistivity increases linearly with temperature, the temperature coefficient of resistivity α is defined as the fractional increase in resistivity per unit increase in temperature.
10. Ohm's law is obeyed by many substances, but it is not a fundamental law of nature. It fails if
- (a) V depends on I non-linearly.
 - (b) the relation between V and I depends on the sign of V for the same absolute value of V .
 - (c) The relation between V and I is non-unique.
- An example of (a) is when ρ increases with I (even if temperature is kept fixed). A rectifier combines features (a) and (b). A thyristor shows all the features (a), (b) and (c).

11. When a source of emf ε is connected to an external resistance R , the voltage V_{ext} across R is given by

$$V_{\text{ext}} = IR = \frac{\varepsilon}{R+r} R$$

where r is the internal resistance of the source.

12. (a) Total resistance R of n resistors connected in *series* is given by

$$R = R_1 + R_2 + \dots + R_n$$

- (b) Total resistance R of n resistors connected in *parallel* is given by

$$\frac{1}{R} = \frac{1}{R_1} + \frac{1}{R_2} + \dots + \frac{1}{R_n}$$

13. Kirchhoff's Rules -

- (a) *First Rule (Junction Rule)*: At any junction of circuit elements, the sum of currents entering the junction must equal the sum of currents leaving it.
- (b) *Second Rule (Loop Rule)*: The algebraic sum of changes in potential around any closed loop must be zero.

14. A *voltmeter* consists of a galvanometer (of resistance R_G) in series with a high resistance R . It is put in parallel to the circuit element across which the voltage is to be measured. Because of its high resistance ($R + R_G \cong R$), it draws a very small current and thus does not disturb the circuit. If the full scale deflection of the galvanometer occurs for a current I_0 and the voltmeter is to have range V_0 , we have

$$V_0 = I_0 (R + R_G) \text{ i.e., } R = \frac{V_0}{I_0} - R_G \cong \frac{V_0}{I_0}$$

15. An *ammeter* consists of a galvanometer (of resistance R_G) and low resistance R in parallel. The effective resistance of the ammeter is

$$\frac{RR_G}{R+R_G} \cong R \quad \text{for } R \ll R_G$$

Because of its very low resistance, the ammeter placed in series in a circuit does not materially change the current in the circuit to be measured. If the full-scale deflection of the galvanometer occurs for current I_0 and the ammeter is to have a range I_{max} , we have

$$R = \frac{I_0 R_G}{I_{\text{max}} + I_0}$$

16. The *potentiometer* is a device to compare potentials. Since the method involves a condition of no current flow, the device can be used to compare emf's of two sources.

17. The *Wheatstone bridge* is an arrangement of four resistances — R_1, R_2, R_3, R_4 as shown in the text. The null-point condition is given by

$$\frac{R_1}{R_2} = \frac{R_3}{R_4}$$

using which the value of one resistance can be determined, knowing the other three resistances.

Physical Quantity	Symbol	Dimensions	Unit	Remark
Electric current	I	[A]	A	SI base unit
Charge	Q	[T A]	C	
Voltage, Electric potential difference	V	[M L ² T ⁻³ A ⁻¹]	V	Work/charge
Electromotive force		[M L ² T ⁻³ A ⁻¹]	V	Work/charge
Resistance	R	[M L ² T ⁻³ A ⁻²]	Ω	$R = V/I$
Resistivity	ρ	[M L ³ T ⁻³ A ⁻²]	Ω m	$R = \rho l/A$
Electrical conductivity	σ	[M ⁻¹ L ³ T ³ A ²]	S	$\sigma = 1/\rho$
Electric field	E	[M L T ⁻³ A ⁻¹]	V m ⁻¹	$\frac{\text{Electric force}}{\text{charge}}$
Drift speed	v_d	[L T ⁻¹]	m s ⁻¹	$v_d = \frac{e E \tau}{m}$
Relaxation time	τ	[T]	s	
Current density	j	[L ⁻² A]	A m ⁻²	current/area
Mobility	μ	[M L ³ T ⁻⁴ A ⁻¹]	m ² V ⁻¹ s ⁻¹	v_d / E

POINTS TO PONDER

- Current is scalar although we represent current with an arrow. Currents do not obey the law of vector addition. That current is a scalar also follows from its definition. The current I through an area of cross-section is given by the scalar product of two vectors:

$$I = \mathbf{j} \cdot \Delta \mathbf{S}$$

where \mathbf{j} and $\Delta \mathbf{S}$ are vectors.

- Refer to V - I curves of a resistor and semiconducting p - n junction diode as drawn in the text. Resistor obeys Ohm's law while a p - n junction diode does not. The assertion that $V = IR$ is a statement of Ohm's law is not true. This equation defines resistance and it may be applied to all conducting devices whether they obey Ohm's law or not. The Ohm's law asserts that the plot of I versus V is linear i.e., R is independent of V . Equation $\mathbf{E} = \rho \mathbf{j}$ leads to another statement of Ohm's law, i.e., a conducting material obeys Ohm's law when the resistivity of the material does not depend on the magnitude and direction of applied electric field.

3. Homogeneous conductors like silver or semiconductors like pure germanium or germanium containing impurities obey Ohm's law within some range of electric field values. If the field becomes too strong, there are departures from Ohm's law in all cases.
4. When we consider conduction in a metal like silver where the conduction electrons move throughout the volume of the metal, we neglect collisions with other electrons but take into account collisions with atoms of the metal. The motion of electron is not governed by classical laws but only by laws of quantum mechanics (why does it happen so? See Chapter 15). Quantum mechanically, the conduction electrons move with temperature-independent speed v_F . For silver, v_F is about 1.6×10^6 m/s.

When an electric field is applied to the metal, the electrons drift slowly in a direction opposite to the field with average drift speed $v_d \sim 5 \times 10^{-7}$ m/s which is much less than v_F by many orders of magnitude.

5. Motion of conduction electrons in electric field \mathbf{E} is the sum of (i) motion due to random collisions and (ii) that due to \mathbf{E} . The motion due to random collisions averages to zero and does not contribute to v_d (Chapter 11, Text book of Class XI). v_d thus is only due to applied electric field on the electron.
6. Electric currents are present in semiconductors in calculators, chips which control microwave ovens and electrical dish washers.

Solar wind is very large electric current of protons, electrons and ions flying radially outward from the Sun. Cosmic rays, by and large, are energetic protons streaming through Milky Way Galaxy.

7. The relation $\mathbf{j} = \rho \mathbf{v}$ should be applied to each type of charge carriers separately. In a conducting wire, the total current and charge density arises from both positive and negative charges:

$$\mathbf{j} = \rho_+ \mathbf{v}_+ + \rho_- \mathbf{v}_-$$

$$\rho = \rho_+ + \rho_-$$

Now in a neutral wire carrying electric current,

$$\rho_+ = -\rho_-$$

Further, $v_+ \sim 0$ which gives

$$\rho = 0$$

$$\mathbf{j} = \rho_- \mathbf{v}_-$$

Thus, the relation $\mathbf{j} = \rho \mathbf{v}$ does not apply to the total current charge density.

8. Kirchhoff's first law is based on conservation of charge and the outgoing currents add up and are equal to incoming current at a junction. Bending or reorienting the wire does not change the validity of Kirchhoff's first law.

EXERCISES

- 3.1** The storage battery of a car has an emf of 12 V. If the internal resistance of the battery is 0.4Ω , what is the maximum current that can be drawn from the battery?
- 3.2** A battery of emf 10 V and internal resistance 3Ω is connected to a resistor. If the current in the circuit is 0.5 A, what is the resistance of the resistor? What is the terminal voltage of the battery when the circuit is closed?
- 3.3** (a) Three resistors 1Ω , 2Ω , and 3Ω are combined in series. What is the total resistance of the combination?
- (b) If the combination is connected to a battery of emf 12 V and negligible internal resistance, obtain the potential drop across each resistor.

- 3.4** (a) Three resistors $2\ \Omega$, $4\ \Omega$ and $5\ \Omega$ are combined in parallel. What is the total resistance of the combination?
 (b) If the combination is connected to a battery of emf 20 V and negligible internal resistance, determine the current through each resistor, and the total current drawn from the battery.
- 3.5** At room temperature ($27.0\ ^\circ\text{C}$) the resistance of a heating element is $100\ \Omega$. What is the temperature of the element if the resistance is found to be $117\ \Omega$, given that the temperature coefficient of the material of the resistor is $1.70 \times 10^{-4}\ ^\circ\text{C}^{-1}$.
- 3.6** A negligibly small current is passed through a wire of length 15 m and uniform cross-section $6.0 \times 10^{-7}\text{ m}^2$, and its resistance is measured to be $5.0\ \Omega$. What is the resistivity of the material at the temperature of the experiment?
- 3.7** A silver wire has a resistance of $2.1\ \Omega$ at $27.5\ ^\circ\text{C}$, and a resistance of $2.7\ \Omega$ at $100\ ^\circ\text{C}$. Determine the temperature coefficient of resistivity of silver.
- 3.8** A heating element using nichrome connected to a 230 V supply draws an initial current of 3.2 A which settles after a few seconds to a steady value of 2.8 A . What is the steady temperature of the heating element if the room temperature is $27.0\ ^\circ\text{C}$? Temperature coefficient of resistance of nichrome averaged over the temperature range involved is $1.70 \times 10^{-4}\ ^\circ\text{C}^{-1}$.
- 3.9** Determine the current in each branch of the network shown in Fig. 3.30:

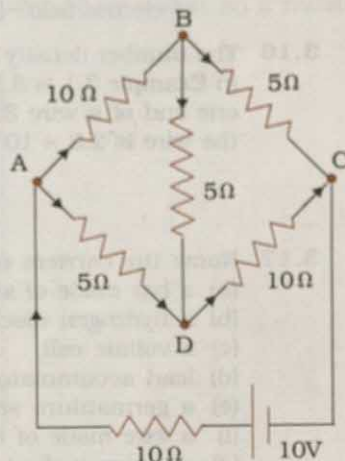


Fig. 3.30

- 3.10** (a) In a metre bridge [Fig. 3.27(a)], the balance point is found to be at 39.5 cm from the end A, when the resistor Y is of $12.5\ \Omega$. Determine the resistance of X. Why are the connections between resistors in a Wheatstone or metre bridge made of thick copper strips?
 (b) Determine the balance point of the bridge above if X and Y are interchanged.
 (c) What happens if the galvanometer and cell are interchanged at the balance point of the bridge? Would the galvanometer show any current?
- 3.11** A storage battery of emf 8.0 V and internal resistance $0.5\ \Omega$ is being charged by a 120 V dc supply using a series resistor of $15.5\ \Omega$. What is the terminal voltage of the battery during charging? What is the purpose of having a series resistor in the charging circuit?
- 3.12** A galvanometer coil has a resistance of $12\ \Omega$ and the metre shows full scale deflection for a current of 3 mA . How will you convert the metre into a voltmeter of range 0 to 18 V ?

- 3.13** A galvanometer coil has a resistance of $15\ \Omega$ and the metre shows full scale deflection for a current of $4\ \text{mA}$. How will you convert the metre into an ammeter of range 0 to $6\ \text{A}$?
- 3.14** In a potentiometer arrangement, a cell of emf $1.25\ \text{V}$ gives a balance point at $35.0\ \text{cm}$ length of the wire. If the cell is replaced by another cell and the balance point shifts to $63.0\ \text{cm}$, what is the emf of the second cell?
- 3.15** Clarify your elementary notions about current in a metallic conductor by answering the following queries:
- In Example 3.1, the electron drift speed is estimated to be only a few mm s^{-1} for currents in the range of a few amperes? How then is current established almost the instant a circuit is closed?
 - The electron drift arises due to the force experienced by electrons in the electric field inside the conductor. But force should cause acceleration. Why then do the electrons acquire a steady average drift speed?
 - If the electron drift speed is so small, and the electron's charge is small, how can we still obtain large amounts of current in a conductor?
 - When electrons drift in a metal from lower to higher potential, does it mean that all the 'free' electrons of the metal are moving in the same direction?
 - Are the paths of electrons straight lines between successive collisions (with the positive ions of the metal) in the (i) absence of electric field, (ii) presence of electric field? (Do check your ideas with the answers at the end).
- 3.16** The number density of conduction electrons in a copper conductor estimated in Example 3.1 is $8.5 \times 10^{28}\ \text{m}^{-3}$. How long does an electron take to drift from one end of a wire $3.0\ \text{m}$ long to its other end? The area of cross-section of the wire is $2.0 \times 10^{-6}\ \text{m}^2$ and it is carrying a current of $3.0\ \text{A}$.

ADDITIONAL EXERCISES

- 3.17** Name the carriers of electric current in
- a bar made of silver
 - a hydrogen discharge tube
 - a voltaic cell
 - lead accumulator being charged by an external supply
 - a germanium semiconductor
 - a wire made of the alloy nichrome
 - a superconductor
- In each case relate the motion of the carrier with the direction of current.
- 3.18** The earth's surface has a negative surface charge density of $10^{-9}\ \text{C m}^{-2}$. The potential difference of $400\ \text{kV}$ between the top of the atmosphere and the surface results (due to the low conductivity of the lower atmosphere) in a current of only $1800\ \text{A}$ over the entire globe. If there were no mechanism of sustaining atmospheric electric field, how much time (roughly) would be required to neutralise the earth's surface? (This never happens in practice because there is a mechanism to replenish electric charges, namely the continual thunderstorms and lightning in different parts of the globe). (Radius of earth = $6.37 \times 10^6\ \text{m}$).
- 3.19**
- Six lead-acid type of secondary cells each of emf $2.0\ \text{V}$ and internal resistance $0.015\ \Omega$ are joined in series to provide a supply to a resistance of $8.5\ \Omega$. What are the current drawn from the supply and its terminal voltage?
 - A secondary cell after long use has an emf of $1.9\ \text{V}$ and a large internal resistance of $380\ \Omega$. What maximum current can be drawn from the cell? Could the cell drive the starting motor of a car?

- 3.20** (a) Make a small survey of the types of primary cells (dry batteries) and secondary cells (accumulators) available in the market and tabulate their important characteristics and uses.
 (b) Besides its emf you might notice another characteristic marked on a storage battery: its 'capacity'. (Do not confuse this term with the capacity of a condenser or capacitor). A storage battery is marked as having a capacity of 3.5 Ah at 1h discharge rate. What does this signify? Would the cell provide 14 A for 15 min?
 (c) Which type of cell would you want to use if your device required
 (i) a current of 100 A for 20 s.
 (ii) a current of 10 mA occasionally?
- 3.21** In a discharge tube, the number of hydrogen ions (i.e. protons) drifting across a cross-section per second is 1.0×10^{18} , while the number of electrons drifting in the opposite direction across another cross-section is 2.7×10^{18} per second. If the supply voltage is 230 V, what is the effective resistance of the tube?
- 3.22** Two wires of equal length, one of aluminium and the other of copper have the same resistance. Which of the two wires is lighter? Hence explain why aluminium wires are preferred for overhead power cables. ($\rho_{\text{Al}} = 2.63 \times 10^{-8} \Omega \text{ m}$, $\rho_{\text{Cu}} = 1.72 \times 10^{-8} \Omega \text{ m}$, Relative density of Al = 2.7, of Cu = 8.9).
- 3.23** What conclusion can you draw from the following observations on a resistor made of alloy manganin:

Current	Voltage	Current	Voltage
A	V	A	V
0.2	3.94	3.0	59.2
0.4	7.87	4.0	78.8
0.6	11.8	5.0	98.6
0.8	15.7	6.0	118.5
1.0	19.7	7.0	138.2
2.0	39.4	8.0	158.0

- 3.24** A cell of emf 1.5 V and internal resistance 0.5Ω is connected to a (non-linear) conductor whose $V-I$ graph is shown in Fig. 3.31. Obtain graphically the current drawn from the cell and its terminal voltage.

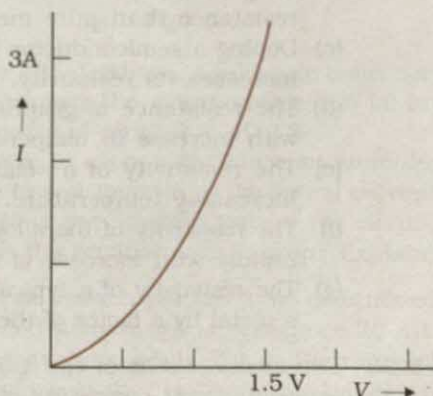


Fig. 3.31

3.25 Answer carefully:

- (a) A simple voltaic cell has an emf equal to 1.0 V. When the circuit is open, is there a net field, which would give rise to a force on a test charge,
 - (i) inside the electrolyte of the cell
 - (ii) outside the cell?
- (b) Is there a net field inside the cell when the circuit is closed and a steady current passes through? Explain.
- (c) Consider a voltaic cell in open circuit with the Cu plate at higher potential with respect to the Zn plate. Since the electrolyte separating them is a conducting medium, why are the positive and negative charges on the electrodes not neutralised immediately? (This question is related to (a) above).
- (d) When a secondary cell of emf 2.0 V is being charged by an external supply, is the terminal voltage of the secondary cell greater or less than 2.0 V?

3.26 Answer the following questions:

- (a) A steady current flows in a metallic conductor of non-uniform cross-section. Which of these quantities is constant along the conductor: current, current density, electric field, drift speed?
- (b) Is Ohm's law universally applicable for all conducting elements? If not, give examples of elements which do not obey Ohm's law.
- (c) It is easier to confine electric currents to definite paths (by the use of electric insulators) than to direct heat flow along definite routes using heat insulators. Why?
- (d) It is easier to start a car engine on a warm day than on a chilly day. Why?
- (e) In which respect does a nearly discharged lead-acid secondary cell differ mainly from a freshly charged cell — in its emf or in its resistance?
- (f) A low voltage supply from which one needs high currents must have very low internal resistance. Why?
- (g) A high tension (HT) supply of, say, 6 kV must have a very large internal resistance. Why?

3.27 Choose the correct alternative:

- (a) Alloys of metals usually have (greater/less) resistivity than that of their constituent metals.
- (b) Alloys usually have much (lower/higher) temperature coefficients of resistance than pure metals.
- (c) Doping a semiconductor (with small traces of impurity atoms) reduces/increases its resistivity.
- (d) The resistance of graphite and most non-metals increases/decreases with increase in temperature.
- (e) The resistivity of a semiconductor increases/decreases rapidly with increasing temperature.
- (f) The resistivity of the alloy manganin is nearly independent of/increases rapidly with increase of temperature.
- (g) The resistivity of a typical insulator (e.g., amber) is greater than that of a metal by a factor of the order of $(10^{22}/10^{23})$.

- 3.28** Two identical cells of emf 1.5 V each joined in parallel provide supply to an external circuit consisting of two resistors of $17\ \Omega$ each joined in parallel. A very high resistance voltmeter reads the terminal voltage of the cells to be 1.4 V. What is the internal resistance of each cell?

- 3.29** (a) Three cells of emf 2.0 V, 1.8 V, and 1.5 V are connected in series. Their internal resistances are $0.05\ \Omega$, $0.7\ \Omega$, and $1\ \Omega$, respectively. If the battery is connected to an external resistor of $4\ \Omega$ via a very low resistance ammeter, what would be the reading in the ammeter?
- (b) If the three cells above were joined in parallel, would they be characterised by a definite emf and internal resistance (independent of external circuit)? If not, how will you obtain currents in different branches of the circuit?
- 3.30** (a) Given n resistors each of resistance R , how will you combine them to get the (i) maximum (ii) minimum effective resistance? What is the ratio of the maximum to minimum resistance?
- (b) Given the resistances of $1\ \Omega$, $2\ \Omega$, $3\ \Omega$, how will be combine them to get an equivalent resistance of (i) $(11/3)\ \Omega$ (ii) $(11/5)\ \Omega$, (iii) $6\ \Omega$, (iv) $(6/11)\ \Omega$?
- (c) Determine the equivalent resistance of networks shown in Fig. 3.32.

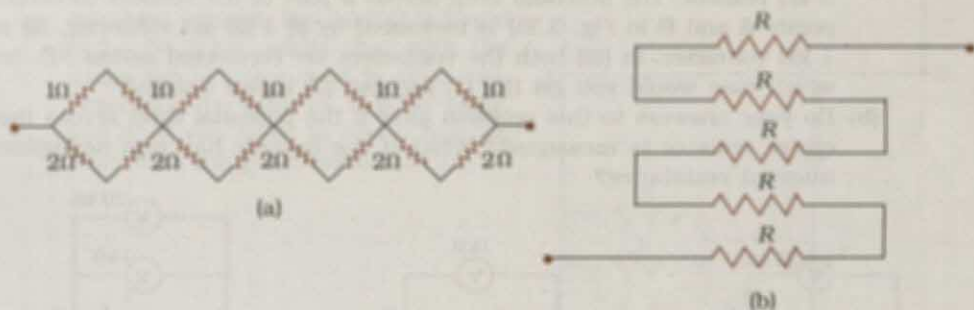


Fig. 3.32(a) and (b)

- 3.31** Determine the current drawn from a 12V supply with internal resistance $0.5\ \Omega$ by the infinite network shown in Fig. 3.33. Each resistor has $1\ \Omega$ resistance.

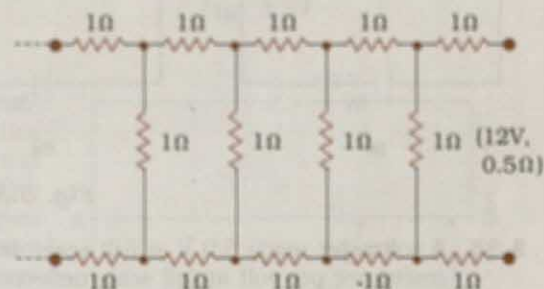


Fig. 3.33

- 3.32** A galvanometer with a coil of resistance $12.0\ \Omega$ shows a full scale deflection for a current of $2.5\ \text{mA}$. How will you convert the galvanometer into (a) an ammeter of range 0 to $7.5\ \text{A}$, (b) a voltmeter of range 0 to $10.0\ \text{V}$. Determine the net resistance of the metre in each case. When an ammeter is put in a circuit, does it read (slightly) less or more than the actual current in the original circuit? When a voltmeter is put across a part of the circuit, does it read (slightly) less or more than the original voltage drop? Explain.
- 3.33** You are given two resistors X and Y whose resistances are to be determined using an ammeter of resistance $0.5\ \Omega$ and a voltmeter of resistance $20\ \text{k}\Omega$. It is known that X is in the range of a few ohms, while Y is in the range of several thousand ohms. In each case, which of the two connections shown in Fig. 3.34 would you choose for resistance measurement? Justify your answer quantitatively. [Hint: For each connection, determine the error in

resistance measurement. The connection that corresponds to a smaller error (for a given range of resistance) is to be preferred.]

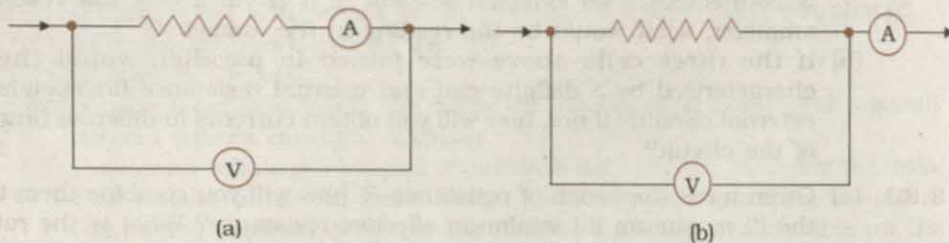


Fig. 3.34(a) and (b)

- 3.34** (a) A battery of emf 9 V and negligible internal resistance is connected to a 3 k Ω resistor. The potential drop across a part of the resistor (between points A and B in Fig. 3.35) is measured by (i) a 20 k Ω voltmeter, (ii) a 1 k Ω voltmeter. In (iii) both the voltmeters are connected across AB. In which case would you get the (1) highest, (2) lowest reading?
- (b) Do your answers to this problem alter if the potential drop across the entire resistor is measured? What if the battery has non-negligible internal resistance?

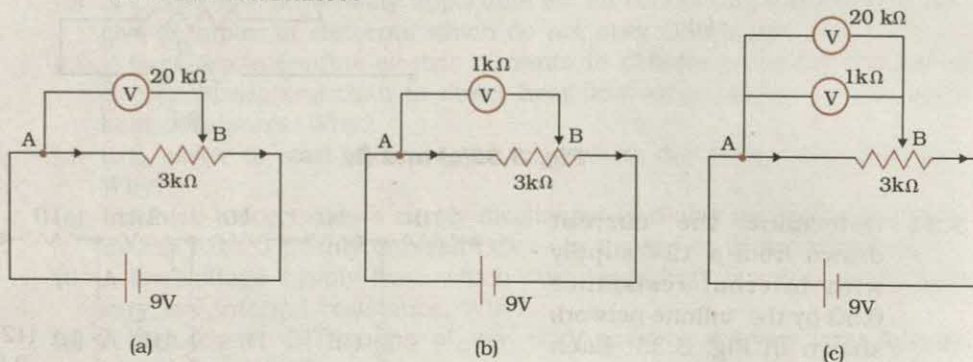


Fig. 3.35

- 3.35** A voltmeter reads 5.0 V at full scale deflection and is graded according to its resistance per volt at full scale deflection as 5000 Ω /V. How will you convert it into a voltmeter that reads 20 V at full scale deflection? Will it still be graded as 5000 Ω /V? Will you prefer this voltmeter to one that is graded as 2000 Ω /V?

- 3.36** Figure 3.36 shows a potentiometer with a cell of 2.0 V and internal resistance 0.40 Ω maintaining a potential drop across the resistor wire AB. A standard cell which maintains a constant emf of 1.02 V (for very moderate currents upto a few mA) gives a balance point at 67.3 cm length of the wire. To ensure very low currents drawn from the standard cell, a very high resistance of 600 k Ω is put in series with it, which is shorted close to the balance point. The standard cell

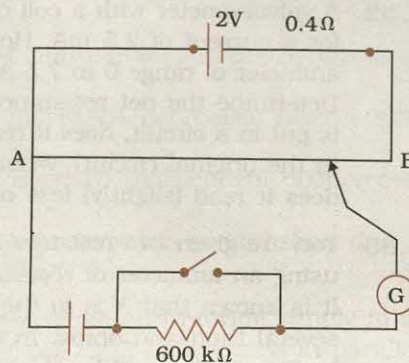


Fig. 3.36

is then replaced by a cell of unknown emf ε and the balance point found similarly, turns out to be at 82.3 cm length of the wire.

- What is the value ε ?
- What purpose does the high resistance of 600 k Ω have?
- Is the balance point affected by this high resistance?
- Is the balance point affected by the internal resistance of the driver cell?
- Would the method work in the above situation if the driver cell of the potentiometer had an emf of 1.0V instead of 2.0V?
- Would the circuit work well for determining an extremely small emf, say of the order of a few mV (such as the typical emf of a thermo-couple)? If not, how will you modify the circuit?

- 3.37** Figure 3.37 shows a potentiometer circuit for comparison of two resistances. The balance point with a standard resistor $R = 10.0 \Omega$ is found to be 58.3 cm, while that with the unknown resistance X is 68.5 cm. Determine the value of X . What might you do if you failed to find a balance point with the given cell of emf ε ?

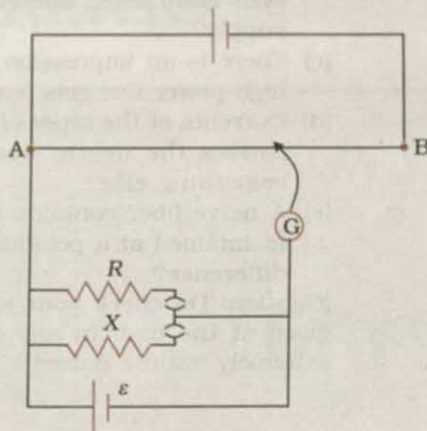


Fig. 3.37

- 3.38** Figure 3.38 shows a 2.0 V potentiometer used for the determination of internal resistance of a 1.5 V cell. The balance point of the cell in open circuit is 76.3 cm. When a resistor of 9.5 Ω is used in the external circuit of the cell, the balance point shifts to 64.8 cm length of the potentiometer wire. Determine the internal resistance of the cell.

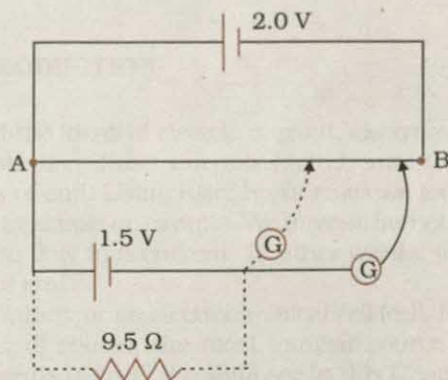


Fig. 3.38

- 3.39** Given a 3.0 V potentiometer, design circuits for calibrating:
- a voltmeter in the range of 0.5 V to 2.5 V
 - an ammeter in the range of 0.1 A to 0.5 A.
- 3.40** A dc supply of 120 V is connected to a large resistance X . A voltmeter of resistance 10 k Ω placed in series in the circuit reads 4 V. What is the value of X ? What do you think is the purpose in using a voltmeter, instead of an ammeter, to determine the large resistance X ?

3.41 Consider a parallel-plate capacitor (of plate area A) during its charging by an external source:

- Show that the conduction current I flowing from (to) the source to (from) the positive (negative) plate equals $\epsilon_0 (dE/dt) A$ where E is the electric field between the plates.
- Is Kirchhoff's first rule valid in this circuit? If not, can you suggest a way to generalize the rule so that it is valid here too?

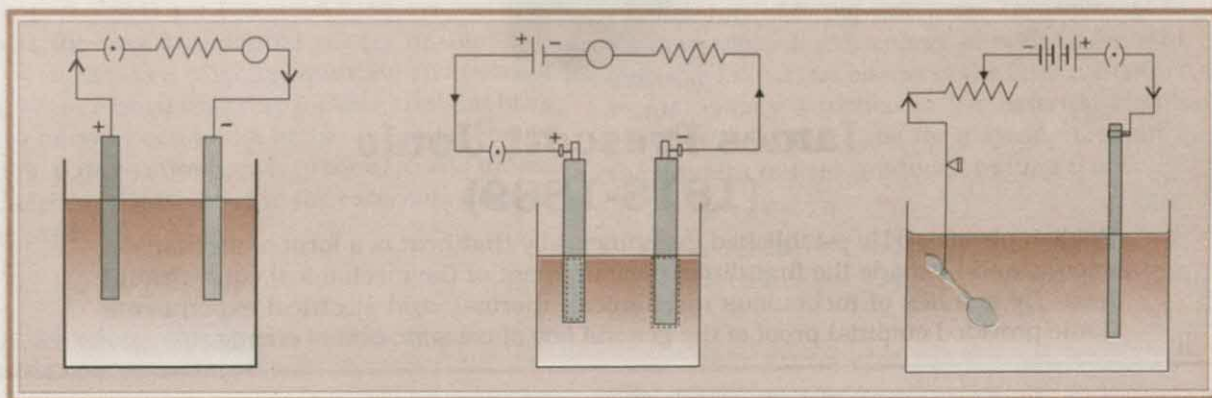
3.42 Answer the following questions:

- What is the order of magnitude of the resistance of a (dry) human body?
- You may be surprised by the answer to (a) above. If the resistance of our body is so large, why does one experience a strong shock (sometimes even fatal) when one accidentally touches the live wire of, say, 240 V supply?
- There is an impression among many people that a person touching a high power line gets 'stuck' with the line. Is that true? Explain.
- Currents of the order of 0.1 A through the human body are fatal. What causes the death: heating of the body due to electric current or something else?
- A nerve fiber contains a membrane separating two conducting 'fluids' maintained at a potential difference. What is the order of this potential difference?

(Caution: Do check your ideas on the above questions with the answers given at the end. In any case, remember that our body is sensitive to extremely minute currents and even small voltages can be risky.)

CHAPTER FOUR

THERMAL AND CHEMICAL EFFECTS OF CURRENTS



4.1 INTRODUCTION

In the preceding Chapter, we have introduced the ideas of electric current, electromotive force (emf), resistance and Ohm's law. We also studied direct current (dc) circuits for different configurations of resistances and the sources of emf. Using Kirchhoff's rules we found that it was possible to obtain distribution of currents in simple dc circuits. We have so far not discussed the question as to what causes the charges to flow in a current. In other words, what is the mechanism that makes something a source of emf?

From practical experience, we know that a battery or an electrochemical cell (cell, for short) is a source of current. The electric generator is, of course, the most familiar source of electric current and power. What makes these things seats of emf? We shall see in this Chapter that in an electrochemical cell, chemical energy is used to produce emf that drives electric current in the circuit. The electric generator, on the other hand, makes use of the principle of electromagnetic induction, and is discussed in Chapter 7. Other sources of emf include photovoltaic cells which involve conversion of solar energy into electrical energy and thermocouples.

In an electrochemical cell, chemical energy is used to drive current. We could also do the reverse: use electric current to effect chemical reactions. Electrolysis is an example of a chemical effect of current. Besides its practical use, the phenomenon of electrolysis had an important role in the history of physics. As you will learn later in this Chapter, one of the first hints on the discrete nature of electric charge came from Faraday's laws of electrolysis.

Thus in a source of emf, chemical or some other form of energy is converted into electrical energy. To maintain the current, the source has to keep expending its energy. Where does this energy go? Part of the source energy in maintaining the current may go into useful work (like in an electric motor). But if the circuit is purely resistive, i.e., it is only a configuration of resistors

and a source of emf, the source energy continually gets dissipated entirely in the form of heat produced in the resistors. This is the heating effect of current that we discuss in the next section. Once again (like in the case of chemical energy) just as current produces thermal energy (heat), thermal energy can be employed suitably to produce a source of emf. The phenomena related to the latter aspect belongs to the subject of thermoelectricity (Section 4.6).



James Prescott Joule **(1818-1889)**

English physicist. He established experimentally that heat is a form of mechanical energy, and he made the first direct measurement of the mechanical equivalent of heat. By a series of meticulous mechanical, thermal, and electrical experiments Joule provided empirical proof of the general law of conservation of energy.



Lord Kelvin (William Thomson) **(1824-1907)**

Born in Belfast, Ireland, is among the foremost British scientists of the nineteenth century. Thomson played a key role in the development of the law of conservation of energy suggested by the work of James Joule (1818-1894), Julius Mayer (1814 - 1878) and Hermann Helmholtz (1821-1894). He collaborated with Joule on the so-called Joule-Thomson effect: cooling of a gas when it expands into vacuum. He introduced the notion of the absolute zero of temperature and proposed the absolute temperature scale, now called the Kelvin scale in his honour. From the work of Sadi Carnot (1796-1832), Thomson arrived at a form of the Second Law of Thermodynamics. Thomson was a versatile physicist, with notable contributions to thermoelectricity, electromagnetic theory and hydrodynamics.

4.2 HEATING EFFECTS: JOULE'S LAW

Consider a steady current I flowing through a circuit. Let the voltage difference between the two ends A and B of the external circuit be V (Fig. 4.1). The current is in the direction of decreasing potential. Suppose Δt is the time for the charge to flow from A to B and let Δq be the quantity of charge that flows across the cross-section at A in time Δt . Then by the definition of current,

$$I = \frac{\Delta q}{\Delta t} \quad (4.1)$$

The potential energies of the charge Δq at the points A and B are $V_A \Delta q$ and $V_B \Delta q$, respectively. Thus the loss in potential energy in time Δt is $(V_A - V_B) \Delta q$, i.e., $V \Delta q$. To maintain the current, the source must make up for this loss and bring the charges back to the higher potential energy. Thus, it must supply energy equal to $V \Delta q$ in time Δt . Hence, power input to the external circuit by the source is

$$P = V \frac{\Delta q}{\Delta t} = VI \quad (4.2)$$

Or, the energy supplied to the external circuit by the source in time t is

$$E = VIt \quad (4.3)$$

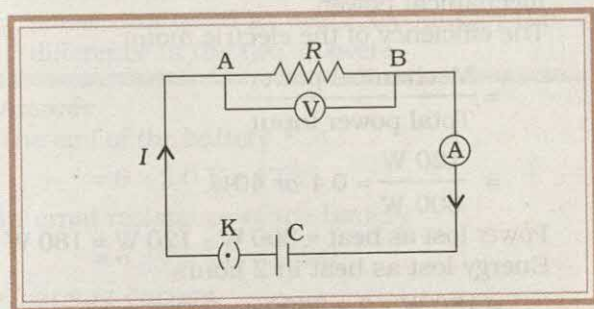


Fig. 4.1 A steady current electric circuit.

What happens to this energy expended by the source? Because of the potential difference between A and B, there is an electric field that acts on the charges. Left to themselves, the charges would accelerate, and the potential energy loss would appear as kinetic energy gain (as happens, for instance, in the free fall of a body). But the situation here is different. The current is steady, i.e., the drift speed of the

charges, on the average, remains constant. Why should this be so, when the charges are being accelerated by the electric field? The reason is that the charge carriers are not isolated. They are continually colliding with the ions or atoms of the lattice of the conducting elements of the circuit. Between two collisions, they pick up kinetic energy due to the field's acceleration. On collision with an ion or atom, there is an exchange of energy. For a steady current, the energy gained by the charge carriers on the whole due to these intermittent accelerations is transferred to the lattice. The lattice ions or atoms receive energy in random bursts at the average rate of IV per unit time. This increase in random (disordered) energy is nothing but the increase in thermal energy of the lattice. In other words, energy supplied to the external circuit appears as heat.* Thus for a steady current I , the amount of heat produced in time t is:

$$H = Pt = VIt \quad (4.4)$$

For a resistive conductor of resistance R , satisfying Ohm's law, Eq. (4.4) can also be written as:

$$H = I^2 R t \quad [4.5(a)]$$

$$= \frac{V^2}{R} t \quad [4.5(b)]$$

Equations [4.5(a)] and [4.5(b)] are known as the **Joule's law of heating**. The law implies that heat produced is (i) directly proportional to the square of current for a given R , (ii) directly proportional to resistance R for a given I , (iii) inversely proportional to resistance R for a given V , and (iv) directly proportional to the time for which the current flows through the resistor.

For alternating current circuits, expressions similar to the above hold good, with the current and voltage replaced by their rms values (Chapter 8).

4.3 PRACTICAL APPLICATIONS OF JOULE'S HEATING

Joule's heating as given by Eq. (4.5) is an inevitable consequence of any electric current. In many cases, it is an undesirable effect, as it converts useful electrical energy into disordered energy (heat). The terms 'ohmic dissipation' or

* This is so if the circuit has no active element like a motor, etc. In general, part of energy supplied by the source goes to do useful work (as in a motor) and the rest appears as heat.

'ohmic loss' are sometimes used to convey the idea of this wasteful conversion of electric energy. For example, in any electric motor, part of the power supplied dissipates as heat. In electric circuits, the unavoidable Joule heating produced in a small region can increase the temperature of the components and alter their properties. Power transmission across long distances is effected at high voltage mainly to reduce current and hence the ohmic loss.

But Joule heating has also many applications. The electric flat iron, bread toaster, oven, electric kettle, electric heater, etc., are familiar domestic applications of Joule heating.

The electric heating may also be used in producing light, as in an electric bulb. Here, the resistor filament must retain as much of the heat generated as possible, so that it gets very hot and emits light. It must not melt at that high temperature. A strong metal with high melting point such as tungsten (melting point 3380°C) is used for bulb filament. The filament should be thermally isolated as much as possible, using insulating support, etc. The bulbs are usually filled with chemically inactive nitrogen and argon gases to prolong the life of filament. Most of the power consumed by the filament appears as heat, but a small part of it is in the form of light radiated. These bulbs give nearly 1 candela light energy for every watt of electrical power consumed.

Another common application of Joule heating is the **fuse** used in electric circuits. It protects a device by preventing any unduly high electric current passing through the circuit. It is placed in series with the device in the circuit and consists of a piece of wire made of a metal of appropriate melting point, e.g., aluminium, copper, iron, lead, etc. If a current larger than the specified value flows through the circuit, the temperature of the fuse wire increases. This melts the fuse wire and breaks the circuit. The fuse wire is usually encased in a cartridge of porcelain or similar material with metal ends. The fuses used for domestic purposes are rated as 1A, 2A, 3A, 5A, 10A, etc. For an electric iron which consumes 1kW electric power when operated at 220 V, a current of $(1000/220)$ A, i.e., 4.54 A will flow in circuit. In this case, a 5 A fuse should be used.

Example 4.1 (a) A small heating element made of nichrome resistor connected to a 30 V dc supply draws a current of 10 A. How much electric power is supplied to the heater? How much heat is produced in 2 hours? (b) An electric motor operating on a 30 V supply draws a current of 10 A and yields mechanical power of 120 W. What is the efficiency of motor? How much energy is lost as heat in 2 hours? (c) Compare and contrast the situations in (a) and (b) above.

Answer

(a) Power supplied to the heater

$$= 30 \text{ V} \times 10 \text{ A} = 300 \text{ W}$$

Since the portion of the circuit to which electric power is supplied is a resistor, the entire electrical energy is converted into heat. Therefore,

heat produced in 2 hours

$$= 300 \text{ W} \times 2 \times 3600 \text{ s} = 2.16 \times 10^6 \text{ J}$$

(b) Power supplied to the motor

$$= 30 \text{ V} \times 10 \text{ A} = 300 \text{ W}$$

The electric motor is not a resistor. Thus, all of the power supplied to it does not appear as heat. Only part of the total power supplied is lost as heat; the rest appears as useful mechanical power.

The efficiency of the electric motor

$$= \frac{\text{Mechanical power}}{\text{Total power input}}$$

$$= \frac{120 \text{ W}}{300 \text{ W}} = 0.4 \text{ or } 40\%.$$

$$= \frac{120 \text{ W}}{300 \text{ W}} = 0.4 \text{ or } 40\%.$$

$$\text{Power lost as heat} = 300 \text{ W} - 120 \text{ W} = 180 \text{ W}$$

$$\text{Energy lost as heat in 2 hours}$$

$$= 180 \text{ W} \times 2 \times 3600 \text{ s} = 1.296 \times 10^6 \text{ J}$$

(c) Electric power supplied in each case is given by VI . The important thing to note is that only for a resistor, $V = IR$ (Ohm's law) and therefore, power supplied is $VI = I^2R$.

For the case (a), $V = IR$ so that the resistance of the heating element is

$$R = 30 \text{ V} / 10 \text{ A} = 3 \Omega, \text{ and } VI = I^2R = 300 \text{ W}.$$

For the case (b), since the electric motor is not a resistor, it is, therefore, wrong to write $VI = I^2R$. For this case, while VI represents power input to the motor, I^2R represents the power lost (as heat) due to the resistance of the element (windings of the motor). The

difference $VI - I^2R$ appears as mechanical power yielded by the motor. To get R for this case, we must equate I^2R to the power lost as heat:

$$I^2R = 180 \text{ W,}$$

i.e.,

$$R = \frac{180}{10 \times 10} = 1.8 \Omega$$

Why is Ohm's law $V = IR$ not directly applicable here? The reason is that when the motor is running, a back emf is developed in the circuit. The current I is then given by (Ohm's law again!)

$$I = (V - V_{\text{back}})/R$$

$$\text{or } VI = V_{\text{back}} I + I^2 R$$

The first term on the right hand side represents the mechanical power output of the electric motor, while the second term represents power loss due to heat.

Example 4.2 A series battery of 6 lead accumulators each of emf 2.0 V and internal resistance 0.25Ω is charged by a 230 V dc mains. To limit the charging current, a series resistance of 53Ω is used in the charging circuit. What is (a) the power supplied by the mains, and (b) the power dissipated as heat? Account for the difference in the two answers.

Answer

The emf of the battery

$$= 6 \times 2.0 \text{ V} = 12 \text{ V}$$

Internal resistance of the battery

$$= 6 \times 0.25 \Omega = 1.5 \Omega$$

Charging current

$$= \frac{(230 - 12) \text{ V}}{(53 + 1.5) \Omega}$$

$$= 4.0 \text{ A}$$

Power supplied by the mains

$$= 230 \text{ V} \times 4.0 \text{ A} = 920 \text{ W.}$$

Power dissipated as heat

$$= 4^2 \times (53 + 1.5) \text{ W} = 872 \text{ W}$$

The difference $(920 - 872) \text{ W} = 48 \text{ W}$, is the power stored in the accumulator in the form of chemical energy of its contents. (Note: You should compare the situation here with that for an

electric motor dealt in example 4.1. The role of 'back emf' there is taken up here by the emf of the battery. The difference between power input and power dissipated equals mechanical power for a motor, while here in the case of lead accumulator, this equals the chemical power stored in the battery).

4.4 CHEMICAL EFFECTS OF CURRENT

4.4.1 Introduction

An important chemical effect of electric current is the phenomenon of electrolysis discussed below. When an electric current passes through a pure metallic conductor, whether solid or liquid (e.g., mercury), there is no chemical effect, i.e., no change in the chemical composition of the conductor takes place. There is only the usual thermal effect (generation of heat). Interesting chemical changes take place when an electric current passes through ionic solutions. A typical experiment as shown in Fig. 4.2 consists of immersing two metal plates or rods in an ionic solution and connecting them to the two ends of a battery. The dissociated ions in the solution appear at the two plates. This is the phenomenon of electrolysis. Quantitative investigations of the phenomenon were first carried out by Michael Faraday. Faraday summarised his findings in the now famous laws of electrolysis. To understand them, let us first look into conduction of electricity in ionic solutions.

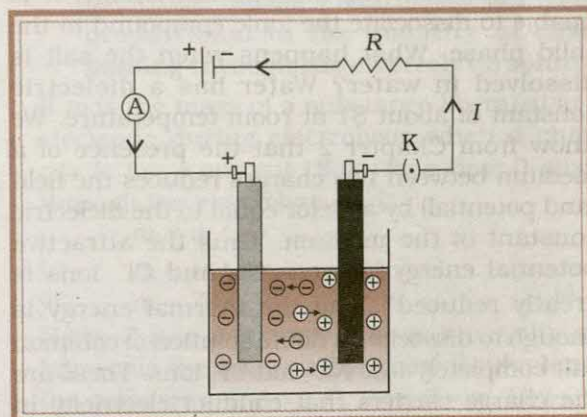


Fig. 4.2 Passage of current through an electrolyte. The positive ions (cations) move towards the cathode and the negative ions (anions) move towards the anode.

4.4.2 Electrolytic Conduction

Electrolytes are substances which conduct electricity because a fraction of their molecules

dissociate into positive and negative ions. The positive ions are called cations and the negative ions are called anions. The conducting pieces (rods, plates or other objects) immersed in the liquid or solid medium and connected to the external circuit are called electrodes. The electrode to which the cations migrate during the conduction of electricity is called the cathode, while that to which the anions migrate is called the anode (Fig. 4.2).

Aqueous solutions of inorganic salts, acids or bases are the most common examples of electrolytes. Many salts such as NaCl, KCl are electrolytes in their molten state, without any solvent. Solutions of organic compounds are, in general, poor conductors of electricity. Pure water does not conduct electricity, but its conductivity increases by a large amount if a few drops of some acid or base are added to it. There are also solid state electrolytes which have mobile ions. Silver Iodide (AgI), for example, has mobile Ag^+ ions.

Let us see how an electrolyte such as common salt or rock salt dissolved in water conducts electricity. Common salt, i.e., crystalline NaCl consists of Na^+ and Cl^- ions bound by electric attraction and arranged as shown in Fig. 4.3. The energy required to dissociate NaCl, i.e. to remove Na^+ and Cl^- ions from each other is about 7.9 eV per molecule. The average thermal energy of 0.03 eV per molecule at room temperature is much less than the dissociation energy; it is unable to dissociate the ionic compound in the solid phase. What happens when the salt is dissolved in water? Water has a dielectric constant of about 81 at room temperature. We know from Chapter 2 that the presence of a medium between two charges reduces the field (and potential) by a factor equal to the dielectric constant of the medium. Thus the attractive potential energy between Na^+ and Cl^- ions is greatly reduced*, and the thermal energy is enough to dissociate a dilute solution of common salt completely into Na^+ and Cl^- ions. These are the charge carriers that conduct electricity in the solution.

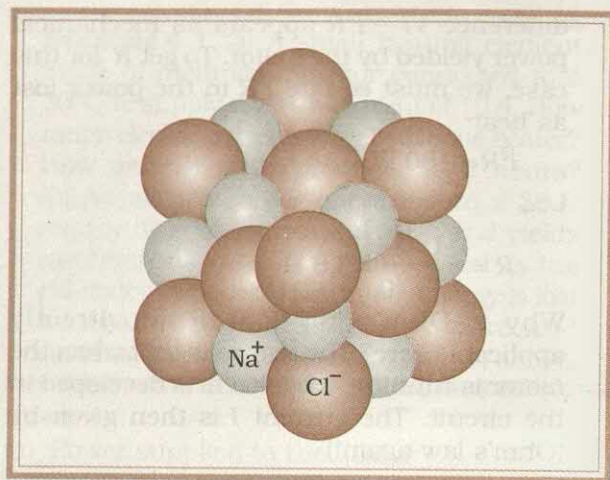


Fig. 4.3 Arrangement of Na^+ and Cl^- ions in a NaCl crystal.

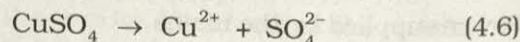
Electrolytes conduct electricity, but their conductivity is much less (by a factor of about 10^{-5} to 10^{-6} at room temperature) than that of metals. The relatively low conductivity arises due to several reasons: smaller number density of ions than that of free electrons, greater viscosity of the medium in which they move, the larger mass of ions and hence, lower drift speed in the external electric field, etc.

4.4.3 Electrolysis

We consider electrodeposition of copper and silver as two simple examples of electrolysis, and then describe the laws of electrolysis, first discovered by Faraday.

Electrolysis of Copper Sulphate Solution

The apparatus in which electrolysis is carried out is called voltameter or electrolytic cell. Two copper plates are partly immersed in aqueous copper sulphate (CuSO_4) solution, which is an electrolyte (Fig. 4.4). The copper plates are connected to the two terminals of a battery. They are the electrodes of the voltameter. The electrolyte is in the form of dissociated copper ions and sulphate ions:



* This is actually an oversimplified picture. The result of Chapter 2 applies if the ions are a large distance apart (compared to their sizes) and the intervening medium (water) can be considered continuous. However, at smaller distances too the attractive energy between the ions is reduced due to the intervening water molecules. Water molecule is polar and the field due to Na^+ and Cl^- ions aligns the dipoles. Further, water molecules around an ion tend to stick to each other due to electrical forces specific to the distribution of electrons in H_2O (hydrogen bonding). Thus, a sizable polarized cluster of water molecules forms around each ion (hydration). The polarisation, as usual, reduces the electric field and the attractive potential energy of the ions. The same reasons account for the large dielectric constant of water.

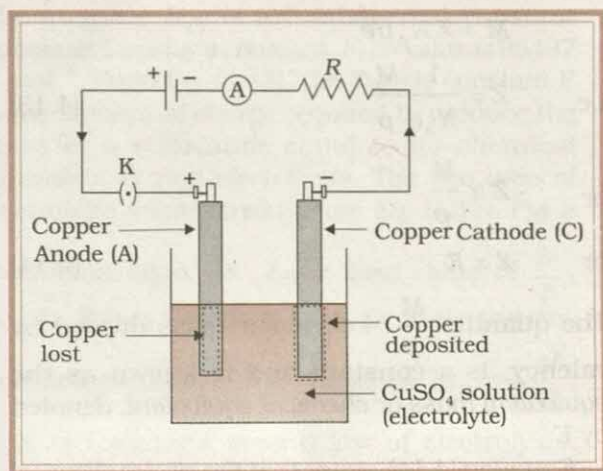


Fig. 4.4 Electrolysis of copper sulphate solution.

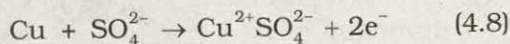
When there is a steady current in the circuit, the following processes take place:

1. Electrons flow from the negative terminal of the battery to the electrode C via the connecting wire.
2. The electrode C is at a lower potential than the electrode A. Therefore, the positively charged copper ions (cations) move towards C, while the negatively charged sulphate ions (anions) move towards A. C is thus the cathode and A the anode of the voltameter.
3. At the cathode C, the Cu^{2+} ions get neutralised by the incoming electrons from the external circuit. The reduction reaction at the cathode is:



The Cu atoms so produced get deposited on the cathode.

4. At the anode, the SO_4^{2-} ions react with copper to release electrons. The oxidation reaction at the anode is:



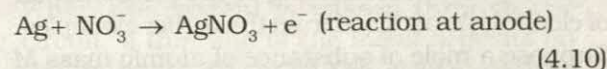
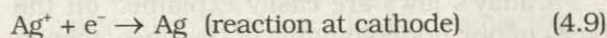
5. The copper ions get into the solution, while the released electrons flow back to the positive terminal of the battery via the connecting metallic wire. The flow of appropriate ions inside the battery then completes the circuit.

Thus, overall, the process results in deposition of copper at the cathode, while the anode loses an equivalent amount of copper.

The concentration of CuSO_4 in the solution remains unchanged. The process is known as *electrodeposition* of copper.

Electrolysis of Silver Nitrate Solution

Here, the voltameter consists of two silver plates partly immersed in the aqueous solution of silver nitrate, which has dissociated in Ag^+ and NO_3^- ions. The description of electrolysis is identical to that for copper sulphate solution, except for one important difference. Silver has valency one, while copper has valency two. The reduction reaction at the cathode and the oxidation reaction at the anode, respectively, are:



As before, silver is deposited at the cathode while the anode loses an equivalent amount of silver. The concentration of AgNO_3 in the solution remains unchanged.

Comparing the two examples, we see that in electrodeposition of silver, one electron circulates for the deposition of one silver atom on the cathode; while in electrodeposition of copper, two electrons circulate for the deposition of one copper atom on the cathode.

4.4.4 Faraday's Laws of Electrolysis

We are now ready to state and understand Faraday's two laws of electrolysis:

1. **The mass of the substance liberated at an electrode during electrolysis is directly proportional to the quantity of charge passing through the electrolytic solution.**

If m is the mass of a substance liberated at an electrode during electrolysis when a charge q ($= It$, i.e., a current I flows for a time t) passes through the electrolyte, then

$$m \propto q$$

$$\text{i.e., } m \propto It$$

$$\text{or } m = ZIt \quad (4.11)$$

Here, Z is a constant of proportionality and known as the **Electro Chemical Equivalent** of the substance. If $q = 1 \text{ C}$, then $Z = m$. Thus, the electrochemical equivalent of a substance is defined as the *mass of the substance liberated when a charge of one coulomb passes through it*.

2. **The masses of different substances liberated by the passage of the same quantity of charge are proportional to their chemical equivalents or equivalent masses.**

Let m_1 and m_2 be the masses of the two substances liberated on the electrodes in two different electrolytic cells when the same current is passed for the same time through their electrolytes. Further, let E_1 and E_2 represent the chemical equivalents of the two substances, respectively. Then,

$$\frac{m_1}{m_2} = \frac{E_1}{E_2} \quad (4.12)$$

Faraday's laws are easily explained on the basis of our understanding of the phenomenon of electrolysis obtained from the earlier examples. Suppose a mole of substance of atomic mass M is deposited on an electrode, i.e., an Avogadro Number N_A of the atoms are deposited on an electrode. If the valency of the atom in the electrolyte is p , then for each atom deposited, a charge ' pe ' must flow through the circuit where e is the electron charge. Therefore, for a mole, the charge flowing through the electrolyte is $N_A pe$. That is, for mass M liberated, the amount of charge that must pass through the circuit is $N_A pe$. From Eq. (4.11), we then have

$$M = Z N_A p e$$

$$\text{i.e.} \quad Z = \frac{1}{N_A e} \frac{M}{p} \quad (4.13)$$

$$\text{or} \quad Z \propto \frac{M}{p}$$

$$\text{or} \quad Z \propto E \quad (4.14)$$

The quantity $\frac{M}{p}$, i.e., molar mass divided by valency, is a constant and is known as the *equivalent mass* or *chemical equivalent*, denoted by E .

Equation (4.14) states that *the electro chemical equivalent (Z) of a substance is directly proportional to its chemical equivalent (or equivalent mass; or mole divided by valency)*. Table 4.1 gives the electro chemical equivalent and chemical equivalent of some substances. The proportionality between Z and E is evident.

Combining Eqs. (4.11) and (4.13), we get

$$m = \frac{M}{N_A e p} It = \frac{E}{F} It \quad (4.15)$$

Table 4.1 Electro Chemical and Chemical Equivalents

Substance	Electro Chemical Equivalent, Z (kg/C)	Atomic Mass (u)	Valency, p	Chemical Equivalent, E (g/mol)
Cations				
Hydrogen	1.045×10^{-8}	1.008	1	1.008
Copper	3.249×10^{-7}	63.57	2	31.78
Silver	1.118×10^{-6}	107.88	1	107.88
Zinc	3.387×10^{-7}	65.39	2	32.695
Chromium	1.800×10^{-7}	51.996	3	17.332
Aluminium	9.360×10^{-8}	27.1	3	9.03
Gold	6.812×10^{-7}	197.2	3	65.73
Nickel	3.040×10^{-7}	58.68	2	29.34
Anions				
Oxygen	8.238×10^{-8}	16	2	8
Chlorine	3.671×10^{-7}	35.46	1	35.46

The quantity $N_A e$ is a fundamental constant known as Faraday's constant, F . Its value is 96487 C mol^{-1} . From Eq. (4.15), the Farady constant F is the amount of charge required to produce the mass of a substance equal to its chemical equivalent during electrolysis. The two laws of electrolysis follow directly from Eq. (4.15). For a

given electrolyte, i.e., for a fixed value of $\frac{M}{p}$, $m \propto It$. This is Faraday's first law of electrolysis.

And for a given amount of charge $q (= It)$, $m \propto \frac{M}{p}$.

This is Faraday's second law of electrolysis. Eq. (4.15) is, therefore, the combined form of Faraday's laws of electrolysis.

Faraday's laws imply that to liberate 1 mole of any substance, a charge equal to Fp is required to pass through the circuit in electrolysis. Since 1 mole contains N_A number of atoms, this means

that the charge per ion of any species is $\frac{Fp}{N_A}$.

This has two important implications:

- (i) The chemical concept of valency is related to electrical charge.
- (ii) There is an elementary charge $e = \frac{F}{N_A}$ and

all charges are multiples of e , since the valency p is an integer.

The laws of electrolysis were thus the first qualitative indicator of quantisation of charge in nature. Interestingly, though this implication was obvious, as great a physicist as Maxwell did not quite reconcile with it and continued to work with continuous charge distributions. (For macroscopic magnitudes of charge, there is not much harm in ignoring quantization of charge, as explained in Chapter 1.) The quantitative significance of the laws is that using the value of F from experiments on electrolysis and the value of N_A from experiments on Brownian motion, the value of elementary charge e can be

calculated ($e = \frac{F}{N_A}$). It turns out that e is about

$1.6 \times 10^{-19} \text{ C}$. This agrees with the value of elementary charge obtained from entirely different experiments such as the Millikan's oil drop experiment (Chapter 12).

4.4.5 Applications of Electrolysis

The phenomenon of electrolysis has many scientific and commercial applications:

1. **Electro-Plating:** The electro-plating of objects by nickel, silver and gold is very common. The conducting material to be electroplated is made the cathode of an electrolytic cell. A strip of metal whose coating is required on the cathode material is used as the anode, while a soluble salt of the same anode material is taken as the electrolyte. Fig. 4.5 shows an experimental set-up used for electro-plating. When the current is passed through the circuit, a thin film of the metal deposits on the cathode. To make the electro-plating uniform and firmly adherent, a suitable current strength is used. If the current strength is very high, the plating may become brittle. For gold plating we need a current from 1 V to 3 V batteries; and for copper, current is drawn from a battery of 5 V to 10 V.

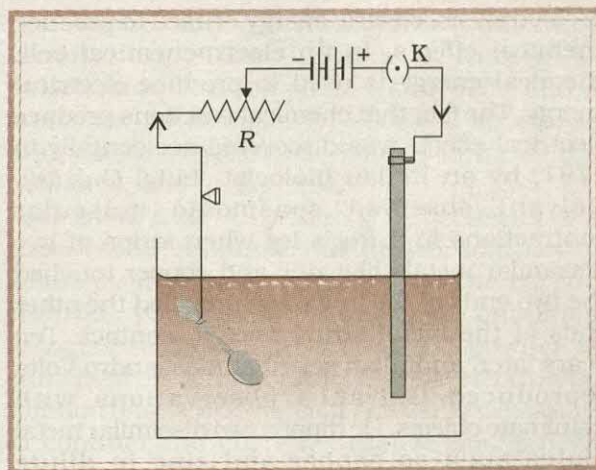


Fig. 4.5 Schematic diagram showing electro-plating.

2. **Extraction of metals from ores:** Certain metals are extracted from their ores using electrolysis. For example, aluminium is obtained by passing an electric current through fused bauxite (Al_2O_3) and cryolite (Na_3AlF_6). Active metals like sodium, calcium and magnesium are also extracted from their ores using electrolysis.
3. **Purification of Metals:** For this purpose, the impure metal is made the anode, and a pure metallic strip is used as cathode. A soluble salt of pure metal is taken as the electrolyte. On passing current, the impure metal anode dissolves but only the pure metal deposits on

the cathode. Many metals like copper are purified up to 99.99% using electrolysis.

4. **Electrolytic Capacitors:** These capacitors consist of two aluminium electrodes placed in an electrolytic mixture of ammonium borate (or sodium phosphate) in glycerine. When a steady current is passed, a thin layer of dielectric aluminium oxide (or hydroxide) is formed on the anode. Such very thin films can offer large values of capacitance. Modern capacitors use electrolytes in the form of a paste or a solution soaked in paper placed between two aluminium foils. If the potential across the two electrodes becomes excessively high, this dielectric layer breaks down and temporarily ceases to function. However, it is possible to regenerate this layer and repair the damage. Such capacitors are very common in power circuits.

4.5 ELECTROCHEMICAL CELLS

In electrolysis, electric energy is used to produce chemical effects. In an electrochemical cell, chemical energy is used to produce electrical energy. The fact that chemical reactions produce electrical effects was discovered accidentally in 1791, by an Italian biologist, Luigi Galvani. Galvani observed spasmodic muscular contractions in a frog's leg when strips of two dissimilar metals like zinc and copper touched the two ends of the frog's leg, provided the other ends of the metal strips were in contact. Ten years later, an Italian scientist Alessandro Volta reproduced Galvani's observations with inanimate objects. He dipped two dissimilar metal plates such as copper and zinc in dilute sulphuric acid and found that a steady current could be produced. This arrangement is called the **Voltaic cell**, an example of an **electrochemical cell** or simply **cell** (Fig. 4.6).

In the Voltaic cell, positive charge accumulates on the copper electrode and negative charge accumulates on the zinc electrode. The copper terminal is, therefore, at a higher potential relative to the zinc terminal. When the terminals are connected externally, the potential difference drives the conventional current from the copper terminal to zinc terminal. Inside the cell, chemical energy drives the current in the opposite direction (from lower to higher potential). Volta observed that the net emf between the two metallic plates depended only on the kind of

metals and conducting electrolyte used and not on the size of the metal plates. He also found that a larger potential difference could be obtained by connecting several cells in series. Such a combination of cells is called a Voltaic battery.

In a chemical reaction, the energy released per mole is typically about 2×10^5 J. This energy moves a mole of ions. If, say, $2e$ is the charge on ions present as reactant or product in the chemical reaction, and V is the emf between the two electrodes of the cell, then

$$2eV N_A = 2 \times 10^5 \text{ J}$$

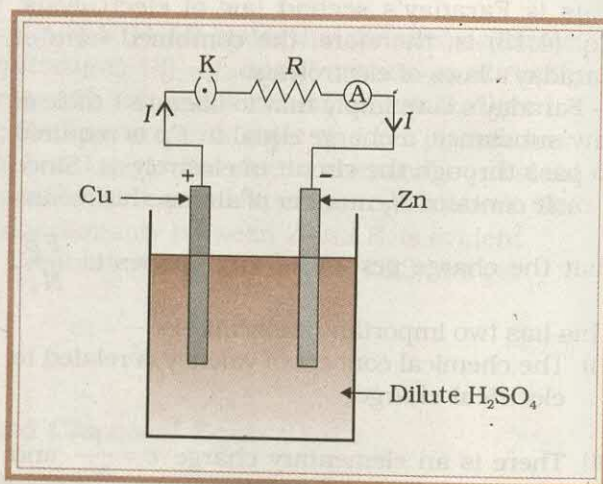


Fig. 4.6 Voltaic cell in an electrical circuit.

This gives V to be about 1 volt. This means that the emf of a cell cannot be very large. Known values of emf are in the range of 1 to 2 V. The total energy a cell can provide is limited by the amount of reactants, unless the reactants can be continuously supplied and the products removed.

Despite the small emf and limited energy capacity, cells or batteries are used in many households and in other applications because of their compactness and convenience. They have no moving parts, are portable and pollution-free. Electronic watches, hearing-aids, car batteries, transistor radios, flash lights, all use cells of various kinds. There is a resurgence of interest in electrochemical cells as source of energy, because many current sources of energy are depleting. We discuss a few types of cells, mostly those in common use.

Daniel Cell: It consists of a copper vessel containing copper sulphate solution. The walls

of copper vessel act as the positive electrode of the cell. A porous pot of fired clay or porcelain containing a zinc rod immersed in dilute sulphuric acid (or acidulated zinc sulphate solution) is placed in the copper sulphate solution (Fig. 4.7). The porous pot allows ions to pass from one solution to the other, but prevents the two solutions to mix. The oxidation reaction takes place at the zinc rod as:



The two electrons are given up to zinc rod and the Zn^{2+} ions move out to the copper sulphate electrolyte through the porous pot. There the Zn^{2+} ions combine with SO_4^{2-} ions to form ZnSO_4 , i.e.,

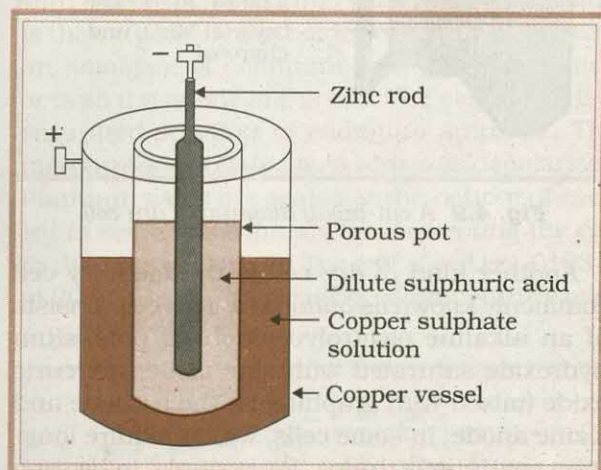
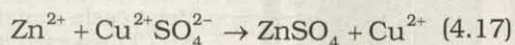
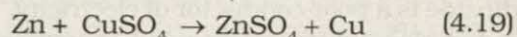


Fig. 4.7 The Daniel Cell.

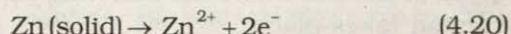
The Cu^{2+} ions move towards the copper vessel and are deposited there. The positive charge is given up to copper vessel in this reduction reaction, i.e.,



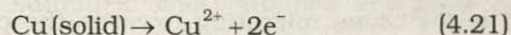
Since inside the cell, positive ion motion is from zinc electrode to copper electrode, the zinc rod is the **anode** and the copper vessel is the **cathode**. If the circuit is completed externally by connecting the zinc rod and copper vessel, electrons flow out of the zinc rod and enter into the copper vessel. The emf of this cell is 1.1 V. In this process two electrons circulate for the loss of every zinc atom from the zinc rod. The energy required per zinc atom is then $2\text{e} \times 1.1\text{ V} = 2.2\text{ eV}$. This is roughly the same as the energy produced per zinc atom in the total reaction:



The emf of the cell can also be taken as the sum of energies released in the half-reactions: oxidation [Eq. (4.16)] and reduction [Eq. (4.18)] reactions at the anode and cathode, respectively. The energies are described in terms of oxidation potentials. The oxidation potentials for the two reactions



and



are 0.76 V and -0.34 V, respectively. Therefore, the emf associated with the full electrochemical reaction [Eq. (4.19)] is $0.76 - (-0.34) = 1.1\text{ V}$.

This cell is not very useful since hydrogen produced by electrolysis in the cell migrates to the copper cathode, where it discharges and bubbles of hydrogen cover the copper vessel. This leads to an increase in the cell's internal resistance and eventually the cell stops functioning. This problem is known as **polarisation**. It is common to many cells, and is overcome, for example, by adding a substance that oxidises hydrogen. This is done in the Leclanche and in dry cells.

Leclanche Cell: A Leclanche cell consists of a glass vessel containing saturated solution of ammonium chloride (electrolyte). The carbon electrode packed in a porous pot containing manganese dioxide and charcoal powder is immersed in ammonium chloride electrolytic solution (Fig. 4.8). Charcoal powder is necessary to conduct the current because manganese

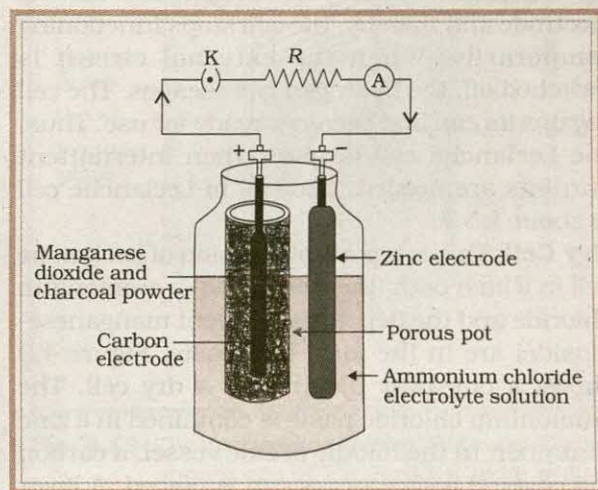


Fig. 4.8 A Leclanche cell.

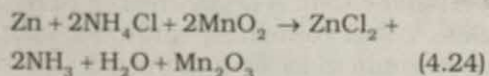
dioxide is a poor conductor of electricity. In the outer glass vessel a zinc rod is immersed in electrolytic solution. When an external circuit is connected across the cell, zinc atoms in contact with the electrolyte ionise, losing two electrons per atom. The electrons flow into the metal wire circuit and Zn^{2+} ions pass into the electrolyte solution. The zinc electrode is the anode of the cell where the following oxidation reaction takes place:



The Zn^{2+} ions migrating towards the carbon cathode combine with Cl^- ions in the electrolyte to form zinc chloride. The released electrons move to the cathode via the external circuit. At the cathode, these electrons combine with NH_4^+ ions, already present in the electrolyte to produce ammonia and hydrogen. The following reduction reaction takes place at the cathode:



The hydrogen reacts with solid manganese dioxide (MnO_2) to form manganese oxide (Mn_2O_3) and water. Thus, MnO_2 prevents hydrogen from collecting on the electrode. The overall chemical reaction in the cell can be summarised as:



The process of absorption of hydrogen (depolarising action) by the solid MnO_2 at the cathode is quite slow. Therefore, if current is drawn from the Leclanche cell continuously, hydrogen gas starts collecting at the carbon electrode and thereby, the cell stops functioning temporarily. When the external circuit is switched off, the hydrogen gas escapes. The cell regains its emf and becomes ready for use. Thus, the Leclanche cell is used when intermittent currents are needed. The emf of Leclanche cell is about 1.5 V.

Dry Cell: This is a portable version of Leclanche cell in which both (the electrolyte — ammonium chloride and the depolarising agent manganese-dioxide) are in the form of a paste. Figure 4.9 shows a cut-away diagram of a dry cell. The ammonium chloride paste is contained in a zinc container. In the middle of zinc vessel, a carbon rod covered with a brass cap is placed. A layer of manganese dioxide and charcoal mixture

surrounds the carbon rod. The whole system is sealed so that the paste does not dry up. However, a small hole is provided near the carbon cathode so that the gas formed during the chemical reaction may escape.

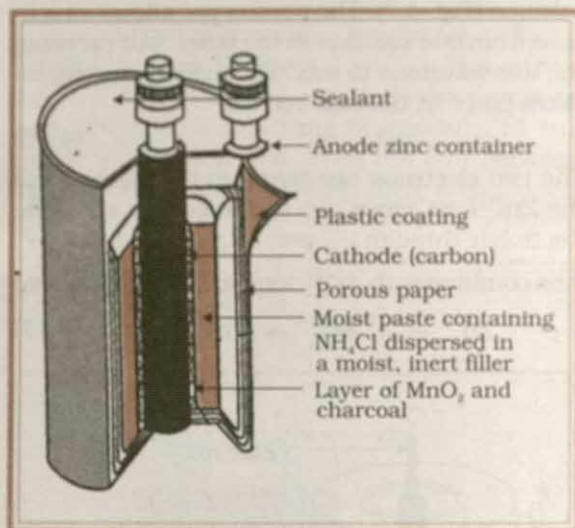


Fig. 4.9 A cut-away diagram of dry cell.

Another kind of dry cell is the mercury cell commonly known as *button cell*. This cell consists of an alkaline electrolyte such as potassium hydroxide saturated with zinc oxide, mercuric oxide (mixed with graphite) as the cathode and a zinc anode. In some cells, which require long-term continuous drains, for example, in hearing aid use, manganese dioxide is added to mercuric oxide in the cathode. Figure 4.10 shows the basic construction of a mercury cell. These cells are very durable, compact, have high energy, density (i.e., amount of electro-chemical energy per unit volume) and offer flat discharge characteristics (constant emf over a large period of time). The emf of a mercury cell is 1.36 V.

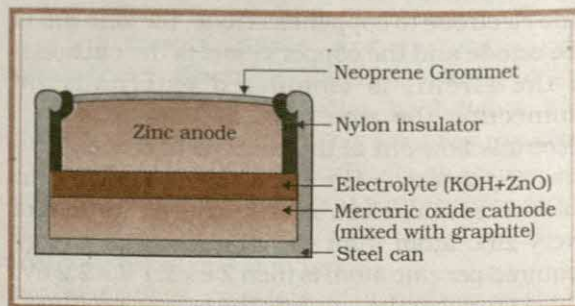


Fig. 4.10 A mercury cell.

Nowadays button cells use silver oxide (mixed with graphite) as the cathode. These silver oxide cells offer 1.62 V emf and have higher energy capacity than the mercury cells. They do not contain any chemical harmful to the environment.

Standard Cell: The electrodes used in the cells discussed so far, deteriorate with the passage of current and cannot offer a constant emf indefinitely. However, there are a few cells called standard cells, which can maintain a fairly constant emf over very long periods of time compared to the other cells. The commonly used standard cell is the Weston Cell (Fig. 4.11). This cell is usually in the form of an H-shaped tube. One leg of the tube contains mercury in contact with a paste of mercurous sulphate (Hg_2SO_4) and is the cathode. The other leg of the tube contains an amalgam of cadmium with mercury, which acts as the anode of the cell. The electrolyte is a saturated solution of cadmium sulphate. The mercurous sulphate paste serves as depolariser. Platinum wires are sealed at the bottom of each leg to serve as terminals for connecting the cell to the external circuit. The emf of cell is 1.0183 V at 20 °C. This is independent of temperature over a considerable range.

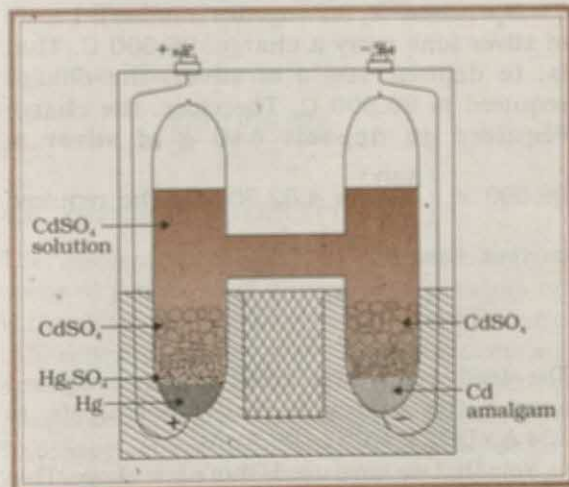
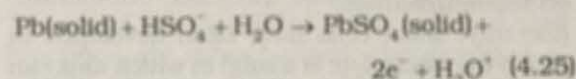


Fig. 4.11 Weston Standard cell.

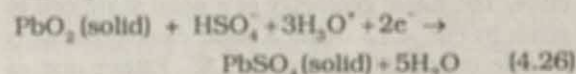
Lead Accumulator: Some cells can be recharged by passing a current through them in the reverse direction. The chemical processes that occur at the electrodes during discharge are then reversed, and the cell recovers its original state, except that

some energy is lost during the charge-discharge cycle. Such cells are called *secondary cells* or *accumulators*. A common example is the lead sulphuric acid cell, invented by Gaston Plante, a French physicist in 1859. The electrodes consist of alternating parallel plates of lead dioxide (positive external electrode or cathode) and spongy lead (Pb, negative electrode or anode) insulated from each other by porous separators made of wood, rubber, plastic or glass fibre. This arrangement is immersed in an electrolyte of dilute sulphuric acid contained in a glass or rubber composition tank (Fig. 4.12).

When the cell operates, the oxidation reaction at the anode is:



The released electrons flow into the cathode via the external circuit where the reduction reaction takes place as:



The overall chemical reaction taking place during the discharge of lead-sulphuric acid cell is:

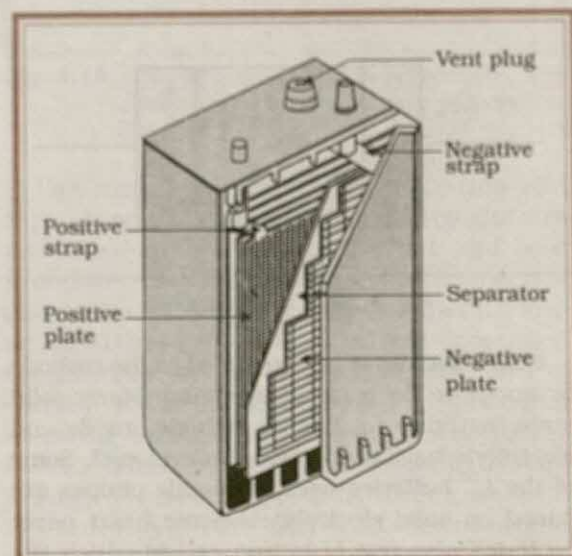
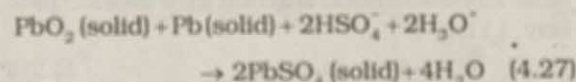


Fig. 4.12 The lead storage battery. The cut-away view shows one of the positive plate grills, a negative plate grill, and separators.

The state of charging is simply monitored by measuring the specific gravity of the electrolyte. It varies from 1.28 when fully charged (sulphuric acid and water) to 1.12 when discharged (mostly water). This secondary cell has a low internal resistance (i.e., it can deliver a high current if needed), and can be recharged a very large number of times without deterioration in properties.

Solid State Cells: The cells discussed above use a 'liquid electrolyte', one cathode and an anode. Some of the major disadvantages of these cells are leakage on long term storage, corrosion due to the use of liquid acidic/alkaline solution, short life, low wattage per kg mass of the cell and limit on miniaturisation.

Recently, some cells have been developed in which the electrolyte is a solid in which ions can move (Solid state electrolytes). Such materials are available in the form of gels, polymers, composites, polycrystalline solids or thin solid films. The basic geometry of a solid state cell is given in Fig. 4.13 using a solid electrolyte with mobile cation M^+ and anion X^- . Either one of these ions or both can move.

In a lithium solid state cell, (Fig. 4.14) the basic electrochemical reaction with the electrode (say, I_2) is

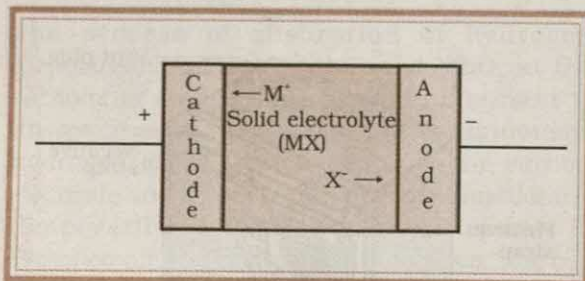
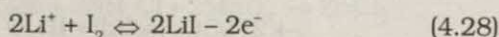


Fig. 4.13 A solid state cell.

Some electrolyte is also mixed in the cathode or anode to decrease polarisation. Many solid state materials for use as cathode, anode and electrolyte have been recently developed. Some of the Li^+ batteries used in mobile phones are based on solid electrolytes. Some heart pacer batteries also use Li-button cell in which the electrolyte is $(LiI + Al_2O_3)$ composite or a similar electrolyte. Polymer Li-batteries and H^+ -batteries

are in advanced stage of development for electrical cars. Some electrode materials like doped $LiCoO_2$ or $LiMnO_2$ have provided excellent rechargeability to these cells.

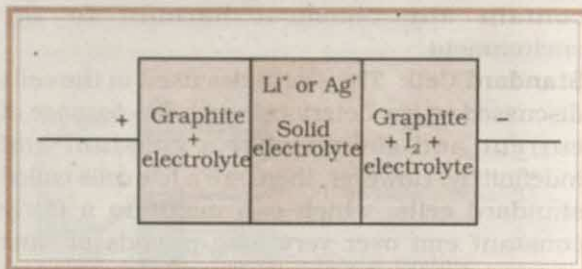


Fig. 4.14 A simple lithium/silver solid state cell.

Example 4.3 It is desired to deposit 0.54 kg of silver per hour on the cathode of a silver voltmeter whose resistance is $0.72 \text{ m}\Omega$. How much potential difference must be maintained between the plates of the voltmeter during electrolysis. Use the known value of Faraday's constant $F = 96,500 \text{ C mol}^{-1}$. Relative atomic mass of silver is 108.

Answer Silver is monovalent. Its ion Ag^+ carries one unit of electronic charge, e . Since $F = N_A e$ (where N_A is Avogadro number), 1 mole of silver ions carry a charge, $96,500 \text{ C}$. That is, to deposit 108 g of silver, the charge required is $96,500 \text{ C}$. Therefore, the charge required to deposit 540 g of silver is

$$96,500 \times \left(\frac{540}{108} \right) = 4,82,500 \text{ C. The required}$$

current, then is

$$I = \frac{4,82,500 \text{ C}}{3,600 \text{ s}} = 134 \text{ A}$$

The steady potential difference required across the plates during electrolysis is $134 \text{ A} \times 0.72 \times 10^{-3} \Omega = 96.5 \text{ mV}$.

Note that we have used Ohm's law above. This is known experimentally to be valid for voltmeters with soluble electrodes (such as silver or copper voltmeters). But for a water voltmeter, for instance, a 'back emf' of about 1.7 V needs to be taken into account before applying Ohm's law.

Example 4.4 A steady current is passed for a certain time through three voltmeters connected in series: a copper voltmeter (Cu electrodes in CuSO_4), a silver voltmeter (Ag electrodes in AgNO_3) and an iron voltmeter (Fe electrodes in FeCl_3). The mass of copper dissolved off the anode of the first voltmeter is found to be 473.1 g. Predict the masses of silver and iron deposited on the respective cathodes of the other two voltmeters. Relative atomic masses of Cu, Ag and Fe are 63.54, 107.9 and 55.85, respectively.

Answer The valencies of Cu (in CuSO_4), Ag (in AgNO_3) and Fe (in FeCl_3) are 2, 1 and 3, respectively. The ratio of relative atomic mass to valency is 31.77 for Cu, 107.9 for Ag, and 18.62 for Fe. According to Faraday's second law of electrolysis, for a given amount of charge (= current \times time) passed, the masses deposited on the cathodes (or dissolved off from the anodes) are in proportion to the ratios of relative atomic mass divided by valency. Therefore, mass of silver deposited

$$m_{\text{Ag}} = \frac{473.1 \times 107.9}{31.77} = 1606.8 \text{ g}$$

and mass of iron deposited

$$m_{\text{Fe}} = \frac{473.1 \times 18.62}{31.77} = 277.2 \text{ g}$$

4.6 THERMOELECTRICITY

Thermoelectricity refers to phenomena that occur at junctions of dissimilar conductors or within a single conductor, when a temperature difference exists between the junctions or across a conductor. We discuss three phenomena: Seebeck effect, Peltier effect and Thomson effect, discovered in that order historically. They involve conversion of thermal energy into electrical energy or vice versa. All the three effects are reversible in contrast to Joule effect which is irreversible conversion of electrical energy into thermal energy. As we shall see, Seebeck effect is a combination of Peltier effect and Thomson effect.

4.6.1 Seebeck Effect

In 1821, Thomson Johann Seebeck, a German physicist, discovered that a current flowed in an electric circuit made of two different conductors, when the two junctions in the circuit were kept at different temperatures. A magnetic compass needle held close to such a circuit showed deflection due to the magnetic field produced by the current. This meant that an emf developed across the two junctions in the circuit causing the flow of current. The two junction circuit is called a thermocouple and the emf developed is called the thermoelectric emf (thermo-emf, for short) or Seebeck emf. Thermo-emf is rather small, generally of the order of μV per degree temperature difference between the junctions. The effect is reversible; if the hot and cold junctions are interchanged, the direction of current reverses. Figure 4.15 shows a circuit made of two dissimilar conductors copper and constantan. The two junctions are maintained at temperatures T_1 and T_2 ($T_1 > T_2$). The direction of current is from copper to constantan at the hotter end (T_1) as shown. If T_2 is the hotter end ($T_1 < T_2$), the current changes sign.

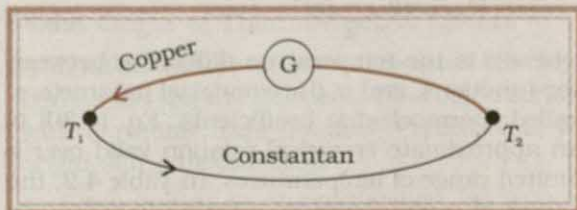


Fig. 4.15 Copper-constantan thermocouple. The direction of current shown corresponds to the case $T_1 > T_2$.

The magnitude and sign of thermo-emf depends on the materials of the two conductors and the temperatures of the hot and cold junctions. Seebeck studied thermoelectric properties of different pairs of metals and arranged the metals in a certain sequence called the **thermoelectric series**. The direction of current, at the hot junction, is from the metal occurring earlier in the series to one occurring later in the series. The magnitude of thermo-emf is larger for metals appearing further apart in the series. A part of Seebeck's thermoelectric series is: Bi, Ni, Co, Pd, Pt, Cu, Mn, Hg, Pb, Sn, Au, Ag, Zn, Cd, Fe, Sb, Te. It should be emphasised that the position of a metal in the

series depends upon the temperature and also on any impurity in the metal.

It is found that emf of a thermocouple AB is the difference between the emfs of two thermocouples AC and BC, provided the junctions are held at the same temperatures:

$$V_{AB} = V_{AC} - V_{BC} \quad (4.29)$$

Thus, we can choose some metal C as standard and determine the thermo-emf of different metals V_{AC} , V_{BC} , etc., for some fixed hot and cold junction temperatures. The thermo-emf of any pair of metals, V_{AB} for the same hot and cold junction temperatures can then be obtained using Eq. (4.29). Lead is usually chosen to be the standard and the reference temperature of cold junction is often taken to be 0°C . For example, at 100°C (cold junction at 0°C), $V_{\text{copper-lead}} = 0.181\text{ mV}$, $V_{\text{constantan-lead}} = -4.255\text{ mV}$. Therefore, $V_{\text{copper-constantan}} = 4.436\text{ mV}$, which is positive giving the direction of current as shown in the Fig. 4.15.

The thermo-emf of a number of thermocouples has the temperature dependance given by the relation:

$$V_{AB} = \alpha\theta + \frac{1}{2}\beta\theta^2 \quad (4.30)$$

where θ is the temperature difference between the junctions, and α, β are material parameters, called thermoelectric coefficients. Eq. (4.30) is an approximate empirical relation valid over a limited range of temperatures. In Table 4.2, the values of α and β are given in the temperature range 200 K to 400 K for a number of substances with lead as the second metal.

Table 4.2
Values of Thermoelectric Coefficients
 α and β .

Substance	α $\mu\text{V}/^\circ\text{C}$	β $\mu\text{V}/(^\circ\text{C})^2$
Aluminium	-0.76	+0.0039
Bismuth (commercial)	-43.7	-0.465
Copper	+1.34	+0.0094
Constantan (60% Cu and 40% Ni)	-38.1	-0.089
Gold	+2.80	+0.010
Iron	+17.2	-0.048
Palladium	-7.4	-0.039
Platinum	-3.04	-0.033
90% Pt and 10% Rh	+7.0	+0.0064

Neutral Temperature and Inversion Temperature

Figure 4.16 shows the temperature variation of thermo-emf of a Cu-Fe thermocouple ($V_{\text{Cu-Fe}}$) with cold junction at 0°C . Note that the thermo-emf rises to a maximum value and then decreases with temperature (of the hot junction). The temperature at which thermo-emf of a thermocouple AB is maximum is called the **neutral temperature** (θ_0); denote as in Eq. (4.31). Mathematically, it is

obtained by finding $\frac{d}{d\theta} V_{AB}$ from Eq. (4.30) and putting it equal to zero, i.e.,

$$\frac{d}{d\theta} V_{AB} = 0,$$

$$\text{or } (\alpha + \beta\theta)|_{\theta=\theta_0} = 0$$

$$\text{i.e., } \theta_0 = -\frac{\alpha}{\beta} \quad (4.31)$$

For copper-iron thermocouple, θ_0 is about 270°C . For most pure metal thermocouples, the neutral temperature is much higher. For thermocouples made of alloys, β is small and thermo-emf varies practically linearly with temperature θ over a wide range of temperature.

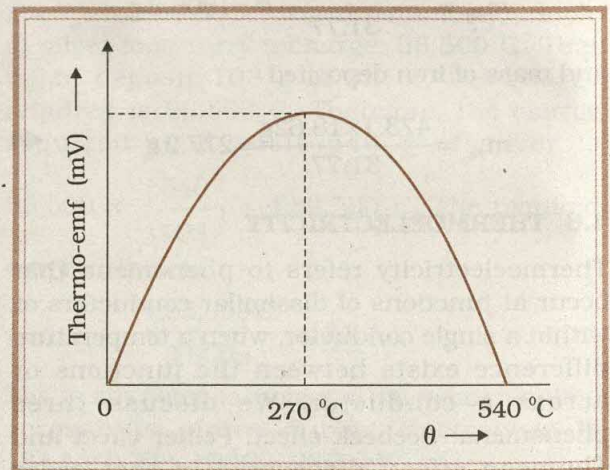


Fig. 4.16 Variation of thermo emf with θ for copper-iron thermocouple.

Beyond the neutral temperature, thermo-emf decreases with temperature and eventually becomes zero at a certain temperature, called the **inversion temperature** θ_i . Beyond the inversion temperature, thermo-emf changes sign and the direction of current reverses.

From Eq. (4.30), putting $V_{AB} = 0$, we get

$$\theta_i = -\frac{2\alpha}{\beta} \quad (4.32)$$

Thus, Eqs. (4.31) and (4.32) show that the *inversion temperature is twice the neutral temperature*, for thermocouples satisfying Eq. (4.30). The reference cold junction is taken to be 0°C .

4.6.2 Peltier Effect

In 1834, a French scientist Peltier, found an effect that was the converse of the Seebeck effect. Figure 4.17 shows a copper-constantan thermocouple again, but now a battery is inserted in the circuit. The direction of the electric current is the same as in Fig. 4.15, i.e., from copper to constantan at the junction 1. It is found that the heat is absorbed at the junction 1 (junction 1 gets cooled) and liberated at the junction 2 (it gets heated). In general, if the Seebeck emf is from A to B at the hot junction, an external emf applied in this direction produces cooling at this junction and heating at the other junction. This phenomenon is known as **Peltier effect**.

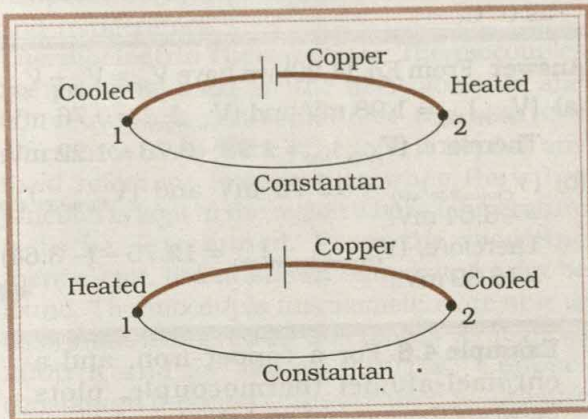


Fig. 4.17 Peltier effect for copper-constantan thermocouple.

As shown in Fig. 4.17, Peltier effect is reversible. Further, the heat absorbed or liberated at a junction is found to be proportional to the first power of current (contrast this with the irreversible Joule effect that is proportional to the square of current.) Thus, if a charge q passes across a junction from metal A to B, heat absorbed at the junction is given by $\pi_{AB}q$ where π_{AB} is known as the Peltier coefficient. The magnitude of π_{AB} depends on the temperature of

the junction. (By this definition, π_{AB} is negative if heat is liberated.) The absorption of heat indicates that there is a seat of emf at the junction, and energy is taken from the environment (resulting in cooling) to provide electrical energy to the current.

4.6.3 Thomson Effect

In 1851, William Thomson (later Lord Kelvin) argued that each element of a single conductor would be a source of emf whenever there is a temperature gradient, i.e., the conductor has a non-uniform temperature. Suppose a current is maintained in a conductor with some temperature gradient. Let ΔT be the temperature difference between the ends of a small section of the conductor. Thomson emf is then given by $\sigma\Delta T$ where σ is called the *Thomson Coefficient*. As before, the existence of emf is indicated by heat absorbed or released per unit quantity of charge transferred. Thomson effect is again reversible. The Thomson coefficient σ depends on temperature and on the material of the conductor.

4.6.4 Origin of Thermoelectric Effects

A detailed theory of thermoelectric effects is beyond the scope of this book. We shall be content with a rudimentary qualitative picture of the origin of thermoelectricity:

1. Each metal is characterised by a work function ϕ – the energy required to remove the highest energy free electron in the metal to infinity (Chapter 12). When two metals A and B are joined together, electrons in the metal with smaller value of ϕ are at a higher energy than those in the metal with greater value of ϕ . Suppose ϕ_A is less than ϕ_B . Clearly, electrons will flow from higher to lower energy levels, i.e., from metal A to metal B. This sets up a potential difference between A and B. The flow continues until the maximum energy levels in the two metals coincide and equilibrium is attained. The potential difference between the metals A and B when equilibrium is attained after diffusion of electrons from A to B is called the *contact potential* between the two metals. The contact potential is proportional to the difference in the work function $\phi_B - \phi_A$.

A still simpler way of explaining the contact potential is to say that electron densities differ from metal to metal. When two metals are in contact, electrons from higher density metal diffuse to the one with lower electron density, thus giving rise to a contact potential between the two metals. This contact potential is the origin of Peltier emf π_{AB} at the junction.

Now, the work function of a metal varies (slightly) with temperature and this variation is different for different metals. Consequently, the contact potential at a junction of two metals varies with temperature. If a closed circuit is made of two dissimilar metals, two junctions are involved. If both are at the same temperature, the two contact potentials are equal and opposite in a closed loop and there is no net emf. If, however, the junctions are held at different temperatures, the contact potentials at the two junctions are different and a net emf will arise in the circuit that will drive a current.

2. In a single conductor with temperature gradient, the free electrons in a region of higher temperature will have higher energy than those in a region of lower temperature. Consequently, there will be a net diffusion of electrons from one region to the other, giving rise to a potential difference. The flow continues until the potential difference is enough to counter the net diffusion due to temperature gradient. This is the origin of Thomson emf. When now a current flows in the conductor, the existence of emf is indicated by absorption or liberation of heat by the conductor.
3. A detailed treatment shows that the Seebeck emf V_{AB} with junctions at temperatures T and T_0 is given by

$$V_{AB} = (\pi_{AB})_T - (\pi_{AB})_{T_0} + (T - T_0)(\sigma_A - \sigma_B) \quad (4.33)$$

Although we omit a rigorous proof of the equation, it is clear qualitatively from the preceding discussion. The Seebeck emf is a result of the difference in Peltier emf at the two junctions and the difference of Thomson emfs as we go along the two conductors in opposite directions in a closed circuit (Fig. 4.18).

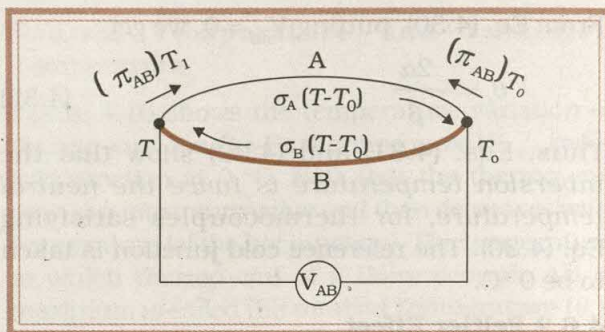


Fig. 4.18 Seebeck, Peltier and Thomson emfs in a thermocouple circuit.

Example 4.5 Find the thermo-emf for the following thermocouple junctions
 (a) Iron-copper thermocouple. Hot junction at 100°C . Thermo-emf of iron-platinum and copper-platinum thermocouples at 100°C are 1.98 mV and 0.76 mV , respectively.
 (b) Chromel-alumel thermocouple. Hot junction at 400°C . Thermo emf of chromel-platinum and alumel-platinum thermocouples at 400°C are 12.75 mV and -3.64 mV , respectively. The temperature of cold junction in both the cases can be taken as 0°C .

Answer From Eq. (4.29) we have $V_{AC} = V_{AB} - V_{CB}$.

(a) $(V_{Fe-Pt})_{100} = 1.98\text{ mV}$ and $(V_{Cu-Pt})_{100} = 0.76\text{ mV}$.

Therefore, $(V_{Fe-Cu})_{100} = 1.98 - 0.76 = 1.22\text{ mV}$.

(b) $(V_{Chromel-Pt})_{400} = 12.75\text{ mV}$ and $(V_{Alumel-Pt})_{400} = -3.64\text{ mV}$.

Therefore, $(V_{Chromel-Alumel})_{400} = 12.75 - (-3.64) = 16.39\text{ mV}$.

Example 4.6 For a copper-iron, and a chromel-alumel thermocouple, plots between thermo-emf and the temperature θ of the hot junction (when the cold junction is at 0°C) are found to satisfy approximately the parabola equation

$$V = \alpha\theta + \frac{1}{2}\beta\theta^2$$

with $\alpha = 14\text{ }\mu\text{V }^\circ\text{C}^{-1}$ and $\beta = -0.04\text{ }\mu\text{V }^\circ\text{C}^{-2}$ for copper-constantan thermocouple and $\alpha = 41\text{ }\mu\text{V }^\circ\text{C}^{-1}$ and $\beta = +0.002\text{ }\mu\text{V }^\circ\text{C}^{-2}$ for chromel-alumel thermocouple. Which of the two thermocouples would you use to measure temperature in the range of about 500°C to 600°C ?

Answer The temperature θ_0 (neutral temperature) corresponding to maximum emf is given by

$$\left(\frac{dV}{d\theta}\right) = 0, \text{ i.e., } \alpha + \beta\theta = 0.$$

which gives $\theta_0 = -(\alpha/\beta)$.

For copper-iron thermocouple, $\theta_0 = 14/0.04 = 350^\circ\text{C}$. (Actually, the coefficients α and β vary slightly with temperature and the neutral temperature for Cu-Fe thermocouple is about 270°C). Beyond the neutral temperature, a thermocouple cannot be used because a given emf corresponds to two different values of θ . Thus, for measuring temperatures beyond about 270°C , a Cu-Fe thermocouple is unsuitable. The chromel-alumel thermocouple, with the given values of α and β , shows no maximum for $\theta > 0$. Thus, it can be used for measuring any high temperature and is suitable for the required range (500°C to 600°C) in the question. In practice, its use is limited to about 1500°C .

4.7 APPLICATIONS OF THERMOELECTRICITY

Thermoelectric effects are mainly used in two kinds of applications in the measurement of temperature and in thermoelectric generators and refrigerators.

Thermoelectric Thermometer: Thermocouples are primarily used for the most accurate and convenient measurement of temperature differences. One of the junctions is kept at any fixed reference temperature while the other junction is kept in the region whose temperature is to be determined. From the resulting thermo-emf, the unknown temperature can be found. Thermocouple thermometers are able to determine temperature differences as high as 2000 K and as low as 0.001 K. Copper-constantan thermocouples are used to measure temperatures in the range 50 K to 500 K. Iron-constantan thermocouples serve up to 1000 K. For higher temperature measurements (up to 2000 K) junctions of platinum-rhodium alloys and junctions of chromel-alumel alloys are convenient. For measurement of very low temperatures (1 K to 50 K), a junction of copper with gold-iron alloy is used because its thermo-emf is large.

Thermoelectric Detector: Thermocouple may also be used as a detector of radiation. If radiant energy in the form of heat or light falls on one junction, it produces a temperature rise and

consequently a thermo-emf is developed in the thermocouple circuit. The small thermoelectric current is measured with a sensitive galvanometer. The sensitivity of thermoelectric detector can be increased if a number of thermocouples are connected in series (Fig. 4.19). Only one set of junctions of the thermocouples is exposed to the heating influence while the other set is protected from radiant energy and kept at room temperature. A device of this sort is called a **thermopile**.

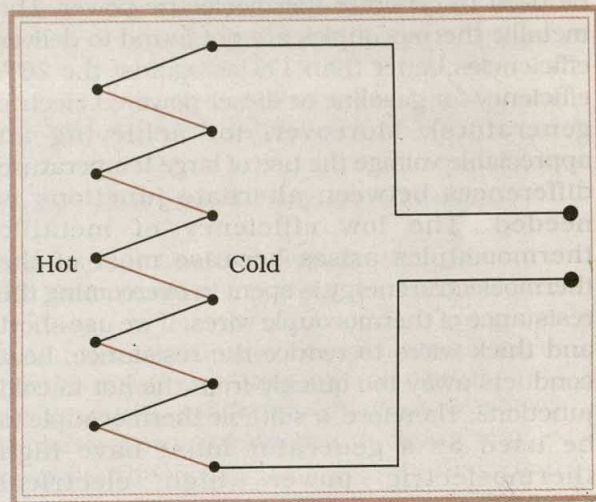


Fig. 4.19 A thermopile.

Thermocouple Current-meter: Thermoelectric effects are also used to measure current. A thermocouple current-meter consists of a resistance R , a thermocouple and a sensitive galvanometer, as shown in Fig. 4.20. The current to be measured passes through the resistor where heat is generated in accordance with I^2Rt heating. This heat warms one junction of the

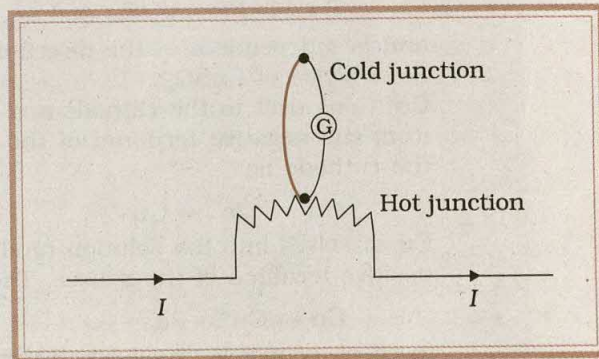


Fig. 4.20 Arrangement of a thermocouple current-meter.

thermocouple while the other junction remains at room temperature. This results in a thermoelectric current that passes through the sensitive galvanometer. The thermoelectric current is roughly proportional to the average rate at which heat is generated, i.e., to the mean value of the square of the alternating current. Thus, the galvanometer reading is roughly proportional to the square of the current being measured.

Thermoelectric Generator: Seebeck effect can be used to generate thermoelectric power. The metallic thermocouples are not found to deliver efficiencies better than 1% (as against the 26% efficiency for gasoline or diesel-powered electric generators). Moreover, for achieving an appreciable voltage the use of large temperature differences between alternate junctions is needed. The low efficiency of metallic thermocouples arises because most of the thermoelectric energy is spent in overcoming the resistance of thermocouple wires. If we use short and thick wires to reduce the resistance, heat conducts away too quickly from the hot to cold junctions. Therefore, a suitable thermocouple to be used as a generator must have high thermoelectric power, high electrical conductivity, and low thermal conductivity. No known metals fulfill these requirements. Semiconductor thermocouples in which the thermal and electrical conductivities can be

controlled separately, offer higher thermoelectric power up to 0.25 mV per K, which is higher than the power offered by metallic thermocouples (up to 0.1 mV per K). Bismuth Telluride (Bi_2Te_3) alloys are promising semiconductor materials for thermoelectric generators. These thermocouples show efficiency of about 7% with nearly 550 K temperature difference between the two junctions. Low power thermoelectric generators have been developed for use in isolated and remote areas. One junction of thermocouple is heated into the shade of kerosene oil lamp while the other junction is kept at room temperature. Recently, using semiconductor alloys, thermoelectric generators have been shown to produce electrical power up to 5 kW.

Thermoelectric Refrigerator: Thermoelectric refrigeration is based on Peltier effect. A junction which on passing current absorbs heat can cool a closed region of space. Such a refrigerator is noiseless, since there are no motors or moving parts. The efficiency of a thermoelectric refrigerator is rather small, but it can be the preferred choice when the region to be cooled is small. For example, with Bi_2Te_3 alloy semiconductor junctions, temperatures as low as 160 K have been achieved. Advances in thermoelectric materials design might dispense with the need for refrigeration using methods that threaten the ozone layer or strengthen the green house effect.

SUMMARY

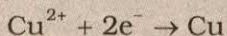
1. If a current I flows through a potential drop V , the energy lost per second by the drifting charges is VI . In a resistor of resistance R , this loss of energy appears as heat. The rate of heat production P is

$$P = VI = I^2R = V^2/R,$$

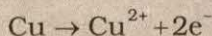
and is independent of the direction of current.

2. Electrolysis of CuSO_4 :

Cu^{2+} ions drift to the cathode and are neutralised by the electrons flowing in from the negative terminal of the external source. The reduction reaction at the cathode is:



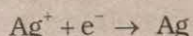
Cu dissolves into the solution producing Cu^{2+} and $2e^-$, the latter flowing to the positive terminal of the source. The oxidation reaction at the anode is:



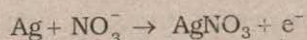
In effect, copper is dissolved off the anode and deposited at the cathode. In this process one copper atom is deposited at the cathode for each pair of electrons flowing through the connecting wires.

3. Electrolysis of AgNO_3 :

Ag^+ ions move towards the cathode and are neutralised by the electron of the metal, flowing in from the negative end of the external source. The reduction reaction at cathode is:



The oxidation reaction at anode is



Ag dissolves into the solution and the released electron flow towards the positive end of the external source. In this process one electron circulates for the deposition of every silver atom on the cathode.

4. Faraday's laws of electrolysis:

- (a) The mass of substance (m) deposited at the cathode (or dissolved off the anode) in electrolysis is proportional to the total charge Q passing through the circuit. If the current I is constant and flows for a time t , $q = It$ and

$$m = ZIt$$

where Z is a constant of proportionality called electro chemical equivalent of the substance.

- (b) The masses of different substances produced in electrolysis by the same quantity of charge are proportional to their equivalent masses (or chemical equivalents).

These laws are easy to understand in the atomic view. For each atom of valency p to be deposited, a charge pe must pass through the circuit. Therefore, for a mole of the substance deposited, i.e., for $m = M$, where M is the relative atomic mass, the total charge flowing through the circuit is $N_A pe$, where N_A is the Avogadro's number. Thus,

$$M = Z N_A pe \text{ or } Z = M/N_A e p$$

where M/p is the equivalent mass of the substance. Thus,

$$m = \frac{Mit}{N_A pe}$$

For a given substance, i.e., for a fixed value of $\frac{M}{p}$, $m \propto It$. This is the first

law of electrolysis. For a given amount of charge It , $m \propto \frac{M}{p}$, which is the second law of electrolysis.

5. Faraday's second law of electrolysis has important implications. It suggests that valency and electrical charge are connected, and that there is an elementary unit of charge e common to all matter so that all charges are integral multiples of e . The fundamental constant $N_A e$ is called Faraday constant (or Faraday) and denoted by F . From the measured values of $F = 96,487 \text{ C/mol}$ (from experiments on electrolysis) and of $N_A = 6.02 \times 10^{23}/\text{mol}$ (from experiments on Brownian motion), e is estimated to be $1.6 \times 10^{-19} \text{ C}$.
6. In *electrochemical cells*, chemical reactions are a source of electrical energy. Examples of such cells are the Daniel cell and the carbon-zinc dry cell. In a secondary cell or accumulator such as the lead sulphuric acid cell, chemical processes that occurred at the electrodes during discharge are reversed by passing a current through it in the reverse direction, and the cell gets recharged, i.e., recovers its original state.
7. (a) *Seebeck effect*: If two junctions of dissimilar conductors in a circuit are held at different temperatures, an emf develops causing a current to flow

in the circuit. The thermo-emf V_{AB} for two materials A and B is the difference between the emfs of two thermocouples AC and BC, provided the junctions are held at the same temperatures, i.e., $V_{AB} = V_{AC} - V_{BC}$. A thermocouple provides an accurate and convenient device for measuring temperature. The copper-constantan thermocouple is very suitable in the range 50 K - 400 K. It develops a voltage of the order of 40 μ V per K. This effect is also used in thermoelectric generators.

- (b) *Peltier effect*: When an electric current is passed through a junction of two dissimilar conductors, heat is either absorbed or released at the junction, depending on the direction of the current. The effect finds application in thermoelectric refrigerators.
- (c) *Thomson effect*: This refers to the emf that develops between two parts of a single metal when they are at different temperatures.

Quantity	Symbol	Dimensions	Units	Remarks
Chemical equivalent	E	$[M \text{ mol}^{-1}]$	kg mol^{-1}	Molar mass/valency
Electrochemical equivalent	Z	$[ML^0T^{-1}A^{-1}]$	kg C^{-1}	$Z = m/It$
Faraday constant	F	$[M^0L^0TA \text{ mol}^{-1}]$	C mol^{-1}	$F = N_A e$
Seebeck emf	V	$[ML^2T^{-3}A^{-1}]$	volt	
Peltier emf	π	$[ML^2T^{-3}A^{-1}]$	volt	
Thomson coefficient	σ	$[ML^2T^{-3}A^{-1}K^{-1}]$	volt K^{-1}	

POINTS TO PONDER

1. A primary cell is a cell in which electrical energy is produced due to chemical reactions taking place in it. Generally these reactions are irreversible. But there are a few cells such as lithium cells, which can be recharged. In a secondary cell, the electrical energy is first stored in the form of chemical energy and when current is drawn from this cell, chemical energy is converted into electrical energy. The chemical reactions are reversible and therefore, these cells can be recharged.
2. When a current is drawn from a battery, i.e. during the discharge of a battery, there is a fall of potential due to the internal resistance of the battery. Therefore, the terminal voltage is reduced and becomes equal to the difference of emf and loss of potential across the internal resistance of battery. On the other hand, when a battery is charged, both its emf and the voltage drop due to internal resistance are in the same direction. Therefore, the potential difference across the terminals of the battery is greater than its emf, when it is being charged.
3. The cooling or heating of a junction in Peltier effect is such that the direction of Seebeck effect can be predicted. If in Fig. 4.15, cooling and heating were reversed, the current due to Seebeck effect and the external current would add up, producing greater cooling or heating, which in turn would produce greater Seebeck emf and so on. Clearly, such a situation grows without bound and is not physically acceptable.
4. To observe Thomson emf across the two ends of a single conductor with some temperature gradient, a current must pass through it.

EXERCISES

- 4.1** A current of 4.0 A flows through a $12\ \Omega$ resistor. What is the rate at which heat energy is produced in the resistor?
- 4.2** An electric motor operates on a 50 V supply and draws a current of 12 A. If the motor yields a mechanical power of 150 W, what is the percentage efficiency of the motor?
- 4.3** A 10 V battery of negligible internal resistance is charged by a 200 V dc supply. If resistance in the charging circuit is $38\ \Omega$, what is the value of the charging current?
- 4.4** A heating element is marked 210 V, 630 W. What is the current drawn by the element when connected to a 210 V dc mains? What is the resistance of the element?
- 4.5** A 10 V storage battery of negligible internal resistance is connected across a $50\ \Omega$ resistor made of alloy manganin. How much heat energy is produced in the resistor in 1 hour? What is the source of this energy?
- 4.6** An electric motor operating on a 50 V dc supply draws a current of 12 A. If the efficiency of the motor is 30%, estimate the resistance of the windings of the motor.
- 4.7** An electric bulb is marked 100 W, 230 V. If the supply voltage drops to 115 V, what is the heat and light energy produced by the bulb in 20 minutes?
- 4.8** The maximum power rating of a $20\ \Omega$ resistor is 2.0 kW. [This is, the maximum power the resistor can dissipate (as heat) without melting or changing in some other undesirable form]. Would you connect this resistor directly across a 300 V dc source of negligible internal resistance? Explain your answer.
- 4.9** Two heaters are marked 200 V, 300 W and 200 V, 600 W. If the heaters are combined in series and the combination connected to a 200 V dc supply, which heater will produce more heat?
- 4.10** An electric power station (100 MW) transmits fixed power to a distant load through long and thin cables. Which of the two modes of transmission would result in lesser power wastage: power transmission at (i) 20,000 V or (ii) 200 V?
- 4.11** Give the direction of thermoelectric current— (i) at the cold junction of copper-bismuth, (ii) at the hot junction of iron-copper and (iii) at the cold junction of platinum-lead.

ADDITIONAL EXERCISES

- 4.12** A dry cell of emf 1.5 V and internal resistance $0.10\ \Omega$ is connected across a resistor in series with a very low resistance ammeter. When the circuit is switched on, the ammeter reading settles to steady value of 20 A. What is the steady (a) rate of chemical energy consumption of the cell, (b) rate of energy dissipation inside the cell, (c) rate of energy dissipation inside the resistor, and (d) power output of the source?
- 4.13** A series battery of 6 lead accumulators each of 2.0 V and internal resistance $0.50\ \Omega$ is charged by a 100 V dc supply. What series resistance should be used in the charging circuit in order to limit the current to 8 A? Using the required resistor, obtain (a) the power supplied by the dc source, (b) the power dissipated as heat, and (c) the chemical energy stored in the battery in 15 min.

- 4.14** (a) A battery of emf ε and internal resistance r is connected across a pure resistive device (e.g., an electric heater or an electric bulb) of resistance R . Show that the power output of the device is maximum when there is a perfect 'matching' between the external resistance and the source resistance (i.e., when $R = r$). Determine this maximum power output.
 (b) What is the power output of the source above if the battery is short-circuited? What is the power dissipation inside the battery in that case?
- 4.15** A 24 V battery of internal resistance $4.0\ \Omega$ is connected to a variable resistor. At what value of the current drawn from the battery is the rate of heat produced in the resistor maximum?
- 4.16** (a) An electric motor runs on a dc source of emf ε and internal resistance r . Show that the power output of the source is maximum when the current drawn by the motor is $\varepsilon/2r$.
 (b) Show that the power output of an electric motor is maximum when the back emf is one half the source emf, provided the resistance of the windings of the motor is negligible.
 (c) Compare and contrast carefully the situation in this exercise with that in Exercise 4.14 above.
- 4.17** Power from a 64 V dc supply goes to charge a battery of 8 lead accumulators each of emf 2.0 V and internal resistance $(1/8)\ \Omega$. The charging current also runs an electric motor placed in series with the battery. If the resistance of windings of the motor is $7.0\ \Omega$ and the steady supply current is 3.5 A, obtain
 (a) the mechanical energy yielded by the motor in 1 hour, and
 (b) the chemical energy stored in the battery during charging in 1 hour.
- 4.18** Two ribbons A and B are given with the following particulars. For a fixed voltage supply, which of the two ribbons gives rise to a greater rate of heat production?

	A	B
Alloy	Constantan	Nichrome
Length (m)	8.456	4.235
Width (mm)	1.0	2.0
Thickness (mm)	0.03	0.06
Temperature coefficient of resistivity ($^{\circ}\text{C}^{-1}$)	Negligible	Negligible
Resistivity ($10^{-7}\ \Omega\text{m}$)	4.9	11

- 4.19** Two wires made of tinned copper having identical cross-section ($=10^{-6}\ \text{m}^2$) and lengths 10 cm and 15 cm are to be used as fuses. Show that the fuses will melt at the same value of current in each case.
- 4.20** A fuse with a circular cross-sectional radius of 0.15 mm blows at 15 A. What should be the radius of cross-section of a fuse made of the same material which will blow at 30 A?
- 4.21** (a) A nichrome heating element across 230 V supply consumes 1.5 kW of power and heats up to a temperature of $750\ ^{\circ}\text{C}$. A tungsten bulb across

the same supply operates at a much higher temperature of 1600 °C, in order to be able to emit light. Does it mean that the tungsten bulb necessarily consumes greater power?

- (b) Which of the two has greater resistance: a 1 kW heater or a 100 W tungsten bulb, both marked 230 V?

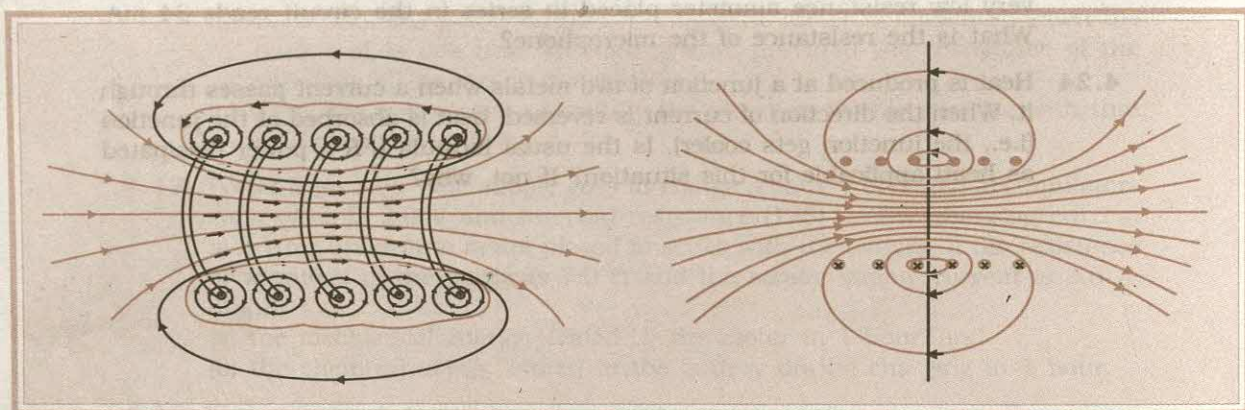
4.22 The thermo-emf ε of a copper constantan thermocouple, and the temperature θ of the hot junction (with cold junction at 0 °C) are found to satisfy approximately the following relation: $\varepsilon = \alpha\theta + \beta\theta^2$ where ε is in μV , θ in °C, and $\alpha = 41 \mu\text{V } ^\circ\text{C}^{-1}$ and $\beta = 0.041 \mu\text{V } ^\circ\text{C}^{-2}$. What is the temperature of the hot junction when the thermo-emf is measured to be 5.5 mV?

4.23 A sensitive microphone cannot withstand currents greater than 0.05 A. When connected across a thermocouple of emf 8.5 mV, the current in a very low resistance ammeter placed in series in the circuit reads 34 mA. What is the resistance of the microphone?

4.24 Heat is produced at a junction of two metals when a current passes through it. When the direction of current is reversed, heat is absorbed at the junction (i.e., the junction gets cooler). Is the usual formula (I^2R = power dissipated as heat) applicable for this situation? If not, why?

CHAPTER FIVE

MOVING CHARGES AND MAGNETISM



5.1 INTRODUCTION

We come across magnets in our daily lives. Magnets are used in audio speakers, motors, transformers and galvanometers. Powerful magnets are used in particle accelerators, magnetic resonance imaging (MRI) machines and in cranes to lift heavy loads. Video and audio recordings as well as computer memories are based on magnetic tapes. In this Chapter, we address a fundamental issue: what is the origin of magnetism?

For a long time it was believed that electricity and magnetism are two distinct phenomena. There was little in common between the two. In the summer of 1820, Hans Christian Oersted made a monumental discovery which demonstrated a close relationship between electricity and magnetism. Before we describe this discovery, we mention that as early as, the seventeenth century, peculiar effects linking electricity to magnetism had been reported. In 1681 a ship headed towards Boston, USA, was struck by lightning. The magnetic compass needles showed wild fluctuations. When the ship was towed to Boston, it was found that some of them had reversed their orientation completely. Presumably, the discharge current associated with the lightning had some effect on them. The invention of the voltaic pile around 1800 (Chapter 4) facilitated many experiments using current electricity. In 1802, an Italian jurist Gian Domenico Romagnosi discovered that a magnetic needle was affected by an electric current flowing in a nearby wire. He even published his observations in a local newspaper. These observations, however, went unnoticed by the scientific community. It took another eighteen years for this effect to be re-discovered, codified, and reported. The credit for this discovery goes to Oersted.

During a lecture demonstration in the summer of 1820, the Danish physicist Hans Christian Oersted noticed that a current in a straight wire caused a noticeable deflection in a nearby magnetic compass needle. He investigated this phenomenon. He found that the alignment of the needle is tangential to an imaginary circle which has the straight wire as its center and has its plane perpendicular to the wire. This situation is depicted in Fig. 5.1(a). It is noticeable when the current is large and the needle sufficiently close to the wire so that the earth's magnetic field may be ignored. Reversing the direction of the current reverses the orientation of the needle [Fig. 5.1(b)]. The deflection increases on increasing the current or bringing the needle closer to the wire. Iron filings sprinkled around the wire arrange themselves in concentric circles with the wire as the center [Fig. 5.1(c)]. Oersted concluded that **moving charges or currents produced a magnetic field in the surrounding space.**

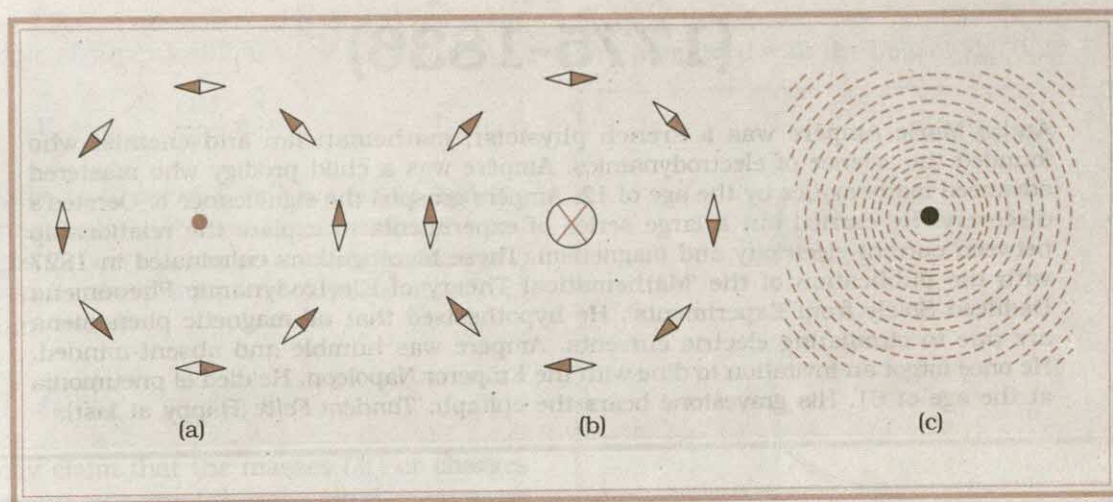


Fig. 5.1 The magnetic field due to a straight long current-carrying wire. The wire is perpendicular to the plane of the paper. A ring of compass needles surrounds the wire. The orientation of the needles is shown when: (a) the current emerges out of the plane of the paper; (b) the current moves into the plane of the paper. (c) The arrangement of iron filings around the wire. The darkened ends of the needle represent north poles. The effect of the earth's magnetic field is neglected.

In this and subsequent Chapters on magnetism, we adopt the following convention: A current or a field (electric or magnetic) emerging out of the plane of the paper is depicted by a dot which is like the tip of a pointed arrow. A current or a field going into the plane of the paper is depicted by a cross which is like the feathered tail of an arrow. Figs. 5.1(a) and 5.1(b) correspond to these two situations, respectively.

Oersted's discovery led to an explosion of scientific activities in the 1820s. The French physicists played a leading role. Francois Arago demonstrated that iron filings are attracted to a current. The current can induce permanent magnetism. This behaviour of a current is similar to an ordinary magnet. Andre-Marie Ampere determined the magnetic field due to a current as also the forces acting between two current-carrying elements. Jean-Baptiste Biot and Felix Savart determined the direction and the distance dependence of a current-carrying element. We describe their epochal discoveries in the following sections.



Andre Ampere (1775-1836)

André Marie Ampère was a French physicist, mathematician and chemist who founded the science of electrodynamics. Ampère was a child prodigy who mastered advanced mathematics by the age of 12. Ampère grasped the significance of Oersted's discovery. He carried out a large series of experiments to explore the relationship between current electricity and magnetism. These investigations culminated in 1827 with the publication of the 'Mathematical Theory of Electrodynamical Phenomena Deduced Solely from Experiments'. He hypothesised that *all* magnetic phenomena are due to circulating electric currents. Ampère was humble and absent-minded. He once forgot an invitation to dine with the Emperor Napoleon. He died of pneumonia at the age of 61. His gravestone bears the epitaph: *Tandem Felix* (Happy at last).



Hans Christian Oersted (1777-1851)

Danish physicist and chemist, professor at Copenhagen. He observed that a compass needle suffers a deflection when placed near a wire carrying an electric current. This discovery gave the first empirical evidence of a connection between electric and magnetic phenomena.

5.2 THE LAW OF BIOT AND SAVART

5.2.1 Sources and Fields

We have seen earlier that the gravitational force \mathbf{F}_1 due to a point mass M_1 on a test point mass m is given very simply by Newton's universal law of gravitation,

$$\mathbf{F}_1 = - \left[\frac{GM_1}{r_1^2} \hat{\mathbf{r}}_1 \right] m$$

where \mathbf{r}_1 is the position vector from M_1 to m and $\hat{\mathbf{r}}_1$ is the corresponding unit vector. The force due to an assembly of point masses (M_i ; $i = 1, 2, \dots, n$) on m is readily stated using the principle of superposition,

$$\mathbf{F}_g = - \left[\sum_{i=1}^n \frac{GM_i}{r_i^2} \hat{\mathbf{r}}_i \right] m \quad (5.1)$$

Similarly, in Chapter 1 we have learnt that the electrostatic force \mathbf{F}_e due to an assembly of point charges (Q_i ; $i = 1, 2, \dots, n$) on a test point charge q is given by the principle of superposition and Coulomb's law,

$$\mathbf{F}_e = \left[\sum_{i=1}^n \frac{Q_i \hat{\mathbf{r}}_i}{4\pi\epsilon_0 r_i^2} \right] q \quad (5.2)$$

We may claim that the masses (M_i) or charges (Q_i) are the *sources* of the force on the test mass m and test charge q , respectively. Alternatively, we can define the quantity in square brackets in Eqs. (5.1) or (5.2) as the field. To be specific, in Chapter 1 we have defined the quantity in square bracket in Eq. (5.2) as the electric field \mathbf{E} . The force on the test charge q is then

$$\mathbf{F}_e = \mathbf{E} q \quad (5.3)$$

We may describe events in terms of *fields* rather than *sources*. If we specify the field, we can specify the force, regardless of the sources.

In magnetism we adopt the 'field approach'. One reason for this is that the sources of magnetic field are not simple. There are no point magnetic charges or monopoles. The known sources are, as we shall shortly see, moving charges or electric currents. We shall see that the magnetic forces are not described in ways analogous to Newton's law of gravitation or Coulomb's law. Contrary to our discussion of gravitation or

electrostatics, we shall first discuss fields and then forces.

5.2.2 The Biot-Savart Law

Soon after Oersted's discovery, the two French physicists Jean-Baptiste Biot and Felix Savart carried out a series of measurements on conductors carrying current and possessing simple shapes. They measured forces between conductors as well as those on a magnetic compass needle placed nearby. From these measurements they correctly guessed the magnetic field $d\mathbf{B}$ due to an elemental conductor $d\mathbf{l}$ carrying current i . The law which relates these is called **Biot-Savart law*** and its statement can be understood with the help of Fig. 5.2.

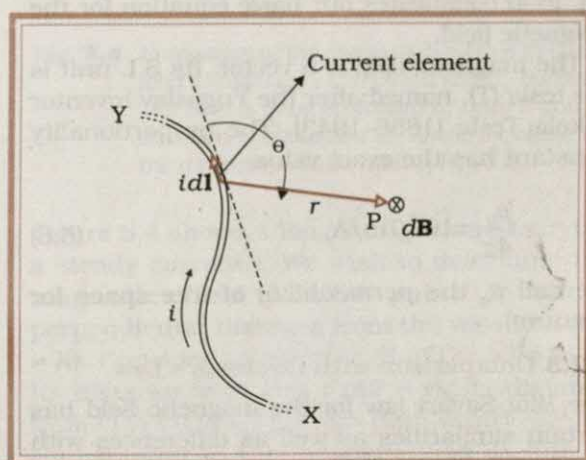


Fig. 5.2 Illustration of the Biot-Savart law. The current element $i d\mathbf{l}$ produces a field $d\mathbf{B}$ at a distance r . The \otimes sign indicates that the field is perpendicular to the plane of this page and directed into it.

Figure 5.2 shows a finite conductor XY carrying current i . Consider an infinitesimal element $d\mathbf{l}$ of the conductor. The field due to this element $i d\mathbf{l}$ is to be determined at a point P which is at a distance r from $i d\mathbf{l}$. Let θ be the angle between $d\mathbf{l}$ and the displacement vector \mathbf{r} . According to Biot and Savart's law, the magnitude of the magnetic field $d\mathbf{B}$ is proportional to the current i , the element length $d\mathbf{l}$, and inversely proportional to the square of the distance r . Its direction is perpendicular to the plane containing $d\mathbf{l}$ and \mathbf{r} . Thus, in vector notation

* Pronounced as "beo and savaar". It rhymes with "rio ad bazaar".

$$d\mathbf{B} \propto i \frac{d\mathbf{l} \times \mathbf{r}}{r^3}$$

$$= \frac{\mu_0}{4\pi} i \frac{d\mathbf{l} \times \mathbf{r}}{r^3} \quad (5.4)$$

where $\mu_0/4\pi$ is a constant of proportionality. The above expression holds when the medium is vacuum.

The magnitude of this field is,

$$|d\mathbf{B}| = \frac{\mu_0}{4\pi} \frac{i dl \sin \theta}{r^2} \quad (5.5)$$

where we have used the property of cross-product

$|d\mathbf{l} \times \mathbf{r}| = dl r \sin \theta$. Just as the square bracket term in Eq. (5.2) defines the electric field, Eq. (5.4) constitutes our basic equation for the magnetic field.

The magnetic field is a vector. Its S.I. unit is the tesla (T), named after the Yugoslav inventor Nikola Tesla (1856-1943). The proportionality constant has the exact value,

$$\frac{\mu_0}{4\pi} = 10^{-7} \text{ Tm/A} \quad (5.6)$$

We call μ_0 the *permeability* of free space (or vacuum).

5.2.3 Comparison with Coulomb's Law

The Biot-Savart law for the magnetic field has certain similarities as well as differences with the Coulomb's law for electrostatic field. Some of these are:

- Both are long range, since both depend inversely on the square of the distance from the source to the point of interest.
- The magnetic field is produced by a vector source $i d\mathbf{l}$. The electrostatic field is produced by a scalar source, namely, the electric charge.
- The electrostatic field is along the displacement vector joining the source and the field point. The magnetic field is perpendicular to the plane containing the displacement vector \mathbf{r} and the current element $i d\mathbf{l}$.
- The principle of superposition applies to both fields. This is because the magnetic field is linearly related to the source $i d\mathbf{l}$ and the electrostatic field is linearly related to its source: the electric charge.

- There is an angle dependence in the Biot-Savart law which is not present in the electrostatic case. In Fig. 5.2, the magnetic field at any point in the direction of $d\mathbf{l}$ (the dashed line) is zero. Along this line, $\theta = 0$

and from Eq. (5.5), $\sin \theta = 0$ and $|d\mathbf{B}| = 0$.

There is an interesting numerical relation between ϵ_0 , the permittivity of free space; μ_0 , the permeability of free space; and c , the speed of light in vacuum.

$$\epsilon_0 \mu_0 = (4\pi \epsilon_0) \left(\frac{\mu_0}{4\pi} \right)$$

$$= \left(\frac{1}{9 \times 10^9} \right) (10^{-7})$$

$$= \frac{1}{(3 \times 10^8)^2} = \frac{1}{c^2} \quad (5.7)$$

We will discuss this connection further in Chapter 9 on electromagnetic waves. Since the speed of light in vacuum is constant, the product $\mu_0 \epsilon_0$ is fixed both in magnitude and dimension. Choosing either ϵ_0 or μ_0 , fixes the other. In SI units, μ_0 is fixed to be equal to $4\pi \times 10^{-7}$ in magnitude.

Example 5.1 An element $\Delta \mathbf{l} = \Delta x \mathbf{i}$ is placed at the origin and carries a large current $i = 10 \text{ A}$. What is the magnetic field on the y -axis at a distance of 0.5 m . $\Delta x = 1 \text{ cm}$.

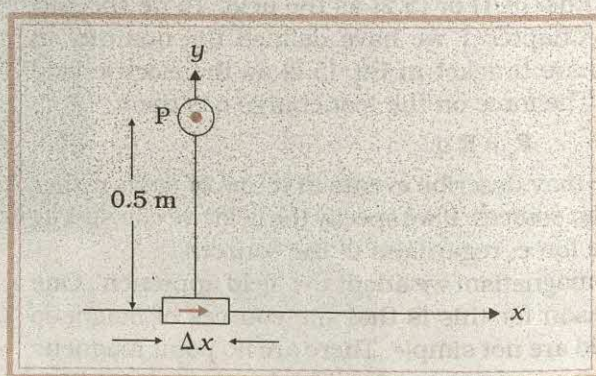


Fig. 5.3 Illustration for Example 5.1.

Answer

$$|d\mathbf{B}| = \frac{\mu_0}{4\pi} \frac{i dl \sin \theta}{r^2} \quad [\text{using Eq. (5.5)}]$$

$$dl = \Delta x = 10^{-2} \text{ m} \quad i = 10 \text{ A}$$

$$r = 0.5 \text{ m} = y \quad \mu_0 / 4\pi = 10^{-7} \frac{\text{Tm}}{\text{A}}$$

$$\theta = 90^\circ; \sin \theta = 1$$

$$|dB| = \frac{10^{-7} \times 10 \times 10^{-2}}{25 \times 10^{-2}}$$

$$= 4 \times 10^{-8} \text{ T}$$

The direction of the field is in the +z-direction. This is so since,

$$d\mathbf{l} \times \mathbf{r} = \Delta x \hat{i} \times y \hat{j}$$

$$= \Delta xy (\hat{i} \times \hat{j})$$

$$= \Delta xy \hat{k}$$

We remind you of the following cyclic property of cross-products,

$$\hat{i} \times \hat{j} = \hat{k}; \hat{j} \times \hat{k} = \hat{i}; \hat{k} \times \hat{i} = \hat{j}$$

Note that the field is small in magnitude. The tesla is a large unit. The highest laboratory fields we can generate are $\leq 50 \text{ T}$. Table 5.1 lists some typical magnetic fields.

In the next section, we shall use the Biot-Savart law to calculate the magnetic field due to conductors of simple shapes and sizes.

Table 5.1 Typical Magnetic Fields

Surface of a neutron star	10^8 T
Large field in a laboratory	1 T
Near a small bar magnet	10^{-2} T
On the earth's surface	10^{-4} T
Human nerve fibre	10^{-10} T
Interstellar space	10^{-12} T

5.3 EVALUATION OF THE MAGNETIC FIELD

In this section, we shall evaluate the magnetic field for two simple configurations: that due to a very long straight wire and that due to a circular coil along its axis. The evaluation entails summing up the effect of miniscule (idl) current elements mentioned in the previous section. In other words, we would be required to carry out an integration. We assume that the current i is steady and that the evaluation is carried out in free space (or vacuum).

5.3.1 Magnetic Field due to a very long Straight Wire

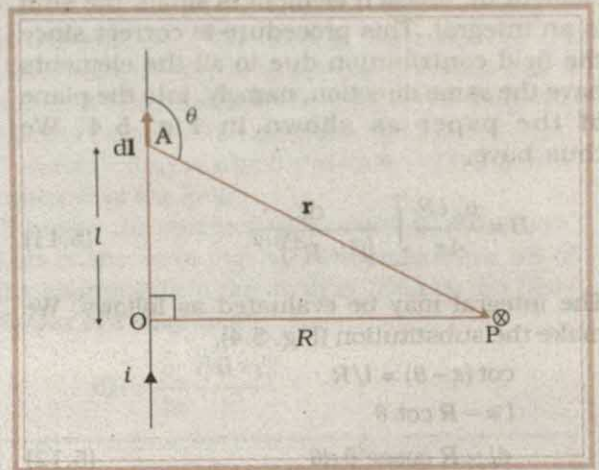


Fig. 5.4 Evaluation of the magnetic field due to a very long straight wire carrying a steady current i . The field is to be evaluated at point P and its direction is shown by \otimes . A rule for obtaining the direction is illustrated in Fig. 5.5.

Figure 5.4 shows a long straight wire carrying a steady current i . We wish to determine the magnetic field at the point P whose perpendicular distance from the wire is R ($OP = R$). Consider an element $d\mathbf{l}$ of the wire at A. Its distance from P is r ($AP = r$). Its distance from O is l ($OA = l$). From Biot-Savart law, its contribution to the magnetic field at P is,

$$d\mathbf{B} = \frac{\mu_0}{4\pi} \frac{i d\mathbf{l} \times \mathbf{r}}{r^3} \quad (5.8)$$

We shall first focus our attention on the magnitude of the magnetic field. From Fig. 5.4 we see that,

$$|d\mathbf{l} \times \mathbf{r}| = dl r \sin \theta \quad (5.9a)$$

and since AOP is a right-angled triangle

$$\sin \theta = \sin(\pi - \theta) = \frac{OP}{AP}$$

$$= \frac{R}{(l^2 + R^2)^{1/2}} \quad (5.9b)$$

where $AP = r = (l^2 + R^2)^{1/2}$. Using Eqs. (5.9a) and (5.9b) in Eq. (5.8) we obtain,

$$dB = \frac{\mu_0}{4\pi} \frac{i R dl}{(l^2 + R^2)^{3/2}} \quad (5.10)$$

In order to obtain the magnetic field due to the entire wire, we simply sum the fields from each element dl . As each element is small, the sum is an integral. This procedure is correct since the field contribution due to all the elements have the same direction, namely, into the plane of the paper as shown in Fig. 5.4. We thus have,

$$B = \frac{\mu_0 i R}{4\pi} \int_{-\infty}^{\infty} \frac{dl}{(l^2 + R^2)^{3/2}} \quad (5.11)$$

The integral may be evaluated as follows. We make the substitution (Fig. 5.4),

$$\begin{aligned} \cot(\pi - \theta) &= l/R \\ l &= -R \cot \theta \\ dl &= R \operatorname{cosec}^2 \theta d\theta \\ l^2 + R^2 &= R^2 \cot^2 \theta + R^2 \\ &= R^2 (\cot^2 \theta + 1) \\ &= R^2 \operatorname{cosec}^2 \theta \end{aligned} \quad (5.12)$$

Thus,

$$\begin{aligned} \frac{dl}{(l^2 + R^2)^{3/2}} &= \frac{R \operatorname{cosec}^2 \theta d\theta}{R^3 \operatorname{cosec}^3 \theta} \\ &= \frac{\sin \theta}{R^2} d\theta \end{aligned} \quad (5.13)$$

We, also need to fix the limits of integration. From Fig. 5.4

$$\tan \theta = -\frac{R}{l}$$

Thus, as

$$l \rightarrow \infty, \theta \rightarrow \pi \quad (5.14a)$$

$$l \rightarrow -\infty, \theta \rightarrow 0 \quad (5.14b)$$

Employing Eqs. (5.12) to (5.14) in Eq. (5.11), we have,

$$\begin{aligned} B &= \frac{\mu_0 i R}{4\pi} \int_0^\pi \frac{\sin \theta}{R^2} d\theta \\ &= \frac{\mu_0 i}{4\pi R} (-\cos \theta) \Big|_0^\pi \\ &= \frac{\mu_0 i}{4\pi R} [\cos(0) - \cos(\pi)] \\ B &= \frac{\mu_0 i}{2\pi R} \end{aligned} \quad (5.16)$$

where we note that $\cos(0) = 1$ and $\cos(\pi) = -1$. The result i.e., Eq. (5.16) is interesting from several points of view. It implies that the field (at every point) on a circle of radius R with the wire as the centre is same in magnitude. The field direction at any point on this circle is tangential to it. Thus the lines of constant magnitude of magnetic field form concentric circles. Notice now, in Fig. 5.1(c) the iron filings form concentric circles. These lines, called **magnetic field lines** form closed loops unlike the electrostatic field lines which originate from and end at charges. What we have done by this exercise is to provide a theoretical justification to Oersted's experiments. Another interesting point to note is that even though the wire is infinite, the field due to it at a finite distance is *not* infinite. It tends to blow up only when we come very close to the wire ($R \rightarrow 0$). The field is directly proportional to the current and inversely proportional to the distance from the (infinitely long) current source.

There exists a simple rule to determine the direction of the magnetic field due to a long wire. This rule, called the **right-hand rule**, is:

Grasp the wire in your right hand with your extended thumb pointing in the direction of the current. Your fingers will curl around in the direction of the magnetic field.

Figure 5.5 illustrate the right-hand rule.

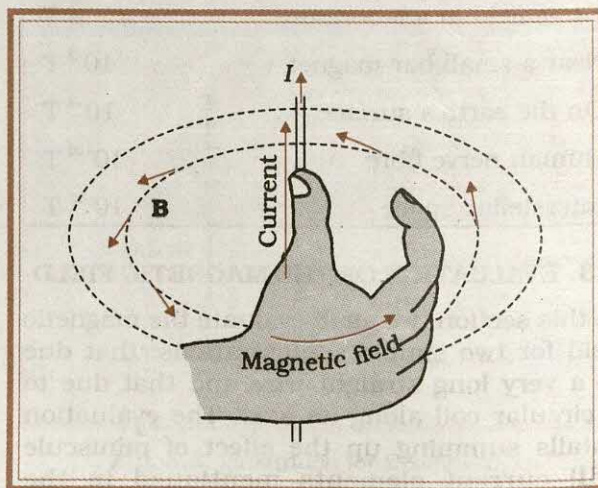


Fig. 5.5 The right-hand rule for the direction of the magnetic field due to a long straight wire. The magnetic field lines form circles with the wire at the centre, reminding us of Fig. 5.1 which illustrates Oersted's experiment.

Example 5.2 Finite wire. Determine the magnetic field due to a finite straight wire carrying a steady current i . The point of interest P makes angles θ_1 and θ_2 with the end-points of the wire as shown in Fig. 5.6.

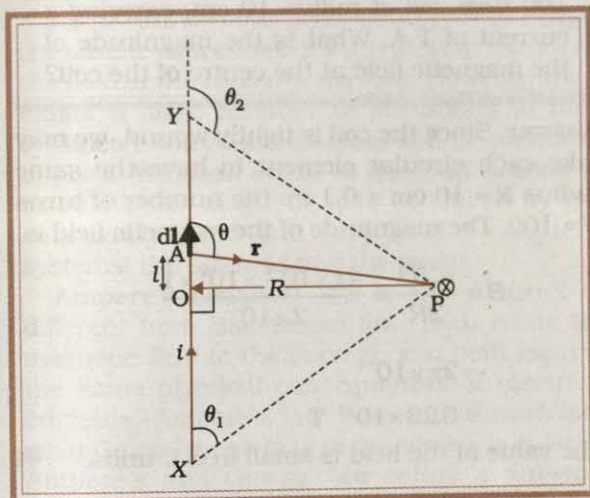


Fig. 5.6 The magnetic field due to a finite wire XY carrying a steady current. The field is to be evaluated at the point P . The lines joining P to the end points of the wire make angles θ_1 and θ_2 with the wire. The evaluation proceeds along the lines very similar to the infinite wire case.

Answer The evaluation of the magnetic field at P (Fig. 5.6) is very similar to the infinite wire case we have just worked out. We borrow the Eq. (5.15) with the limits appropriately calculated from Fig. 5.6, the lower limit is θ_1 and the upper limit is θ_2 .

$$\begin{aligned} B &= \frac{\mu_0 i R}{4\pi} \int_{\theta_1}^{\theta_2} \frac{\sin\theta}{R^2} d\theta \\ &= \frac{\mu_0 i}{4\pi R} (-\cos\theta) \Big|_{\theta_1}^{\theta_2} \\ &= \frac{\mu_0 i}{4\pi R} (\cos\theta_1 - \cos\theta_2) \end{aligned} \quad (5.17)$$

The limiting case of infinite wire can be treated by simply taking $\theta_1 = 0$ and $\theta_2 = \pi$.

5.3.2 Magnetic Field on the Axis of a Circular Current Loop

Figure 5.7 depicts a circular loop carrying a steady current i . The loop is placed in the y - z plane with its centre at the origin O and has a radius R . The x -axis is the axis of the loop. We wish to calculate the magnetic field at the point P on this axis. Let x be the distance of P from the center O of the loop.

Consider the conducting element $d\mathbf{l}$ of the loop. This is shown in Fig. 5.7. The magnitude dB of the magnetic field due to $d\mathbf{l}$ is given by the Biot-Savart law [Eq. (5.4)],

$$dB = \frac{\mu_0}{4\pi} \frac{i |d\mathbf{l} \times \mathbf{r}|}{r^3}$$

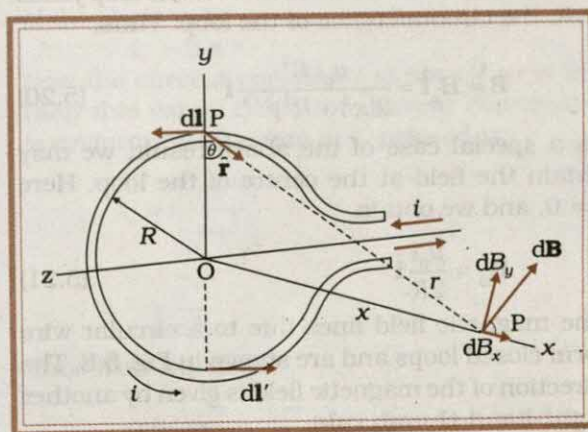


Fig. 5.7 A circular loop of radius R carrying a steady current i .

Now $r^2 = x^2 + R^2$. Further, any element of the loop will be perpendicular to the displacement vector from the loop to the axial point. For example, the element $d\mathbf{l}$ in Fig. 5.7 is in the y - z direction whereas the displacement vector \mathbf{r} from $d\mathbf{l}$ to the axial point P is in the x - y plane. Hence $|d\mathbf{l} \times \mathbf{r}| = r dl$. Thus,

$$dB = \frac{\mu_0}{4\pi} \frac{idl}{(x^2 + R^2)} \quad (5.18)$$

The direction of $d\mathbf{B}$ is shown in Fig. 5.7. It is perpendicular to the plane formed by $d\mathbf{l}$ and \mathbf{r} . It has an x -component dB_x and a y -component dB_y^* . When the components

* You may also argue this algebraically. As shown in Fig. 5.7, $d\mathbf{l} = dl \hat{\mathbf{k}}$ and $\mathbf{r} = x \hat{\mathbf{i}} - R \hat{\mathbf{j}}$. Then $d\mathbf{l} \times \mathbf{r} = dl \times (x \hat{\mathbf{i}} - R \hat{\mathbf{j}}) = dl (R \hat{\mathbf{i}} + x \hat{\mathbf{j}})$ which is a vector in the x - y plane.

perpendicular to the x -axis are summed over, we obtain a null result. For example, the dB_y component due to $d\mathbf{l}$ is cancelled by the contribution due to the diametrically opposite $d\mathbf{l}$ element, shown in Fig. 5.7. Thus, only the x -component survives. The net contribution along x -direction can be obtained by integrating $dB_x = dB \cos \theta$ over the loop. For Fig. 5.7,

$$\cos \theta = \frac{R}{(x^2 + R^2)^{1/2}} \quad (5.19)$$

From Eqs. (5.18) and (5.19),

$$dB_x = \frac{\mu_0 i}{4\pi} \frac{d\mathbf{l} R}{(x^2 + R^2)^{3/2}}$$

The summation of elements $d\mathbf{l}$ over the loop yields $2\pi R$, the circumference of the loop. Thus,

$$\mathbf{B} = B_x \hat{\mathbf{i}} = \frac{\mu_0 i R^2}{2(x^2 + R^2)^{3/2}} \hat{\mathbf{i}} \quad (5.20)$$

As a special case of the above result, we may obtain the field at the centre of the loop. Here $x = 0$, and we obtain,

$$\mathbf{B}_0 = \frac{\mu_0 i}{2R} \hat{\mathbf{i}} \quad (5.21)$$

The magnetic field lines due to a circular wire form closed loops and are shown in Fig. 5.8. The direction of the magnetic field is given by another **right-hand thumb rule**.

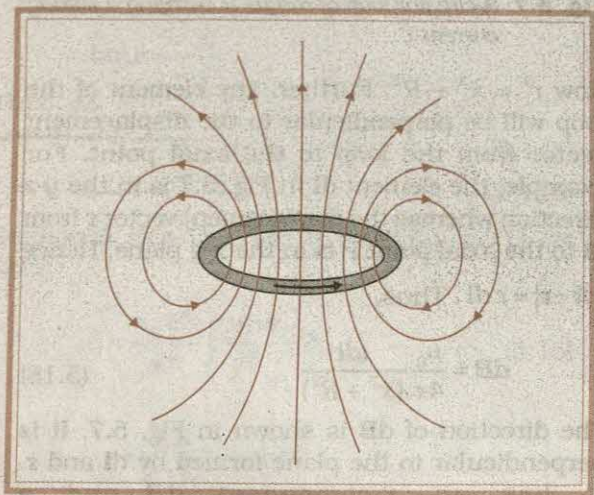


Fig. 5.8 The magnetic field lines for a current loop. The direction of the field is given by the right-hand thumb rule described in the text. The upper side of the loop may be thought of as the north pole and the lower side as the south pole of a magnet.

Curl the palm of your right hand around the circular wire with the fingers pointing in the direction of the current. The right hand thumb gives the direction of the magnetic field.

Example 5.3 Consider a tightly wound 100 turn coil of radius 10 cm, carrying a current of 1 A. What is the magnitude of the magnetic field at the centre of the coil?

Answer Since the coil is tightly wound, we may take each circular element to have the same radius $R = 10 \text{ cm} = 0.1 \text{ m}$, the number of turns $N = 100$. The magnitude of the magnetic field is,

$$\begin{aligned} B &= \frac{\mu_0 N i}{2R} = \frac{4\pi \times 10^{-7} \times 10^2 \times 1}{2 \times 10^{-1}} \\ &= 2\pi \times 10^{-4} \\ &= 6.28 \times 10^{-4} \text{ T} \end{aligned}$$

The value of the field is small in S.I. units. ◀

5.4 AMPERE'S CIRCUITAL LAW

There is an alternative and appealing way in which the Biot-Savart law may be expressed. We shall explain it with the help of Section 5.3.1, where the magnetic field due to a long wire is discussed. Equation (5.16) for the magnitude of this field can be re-expressed as,

$$B(2\pi R) = \mu_0 i \quad (5.22)$$

The left hand side is the product of the magnetic field and the circumference of the circle of radius R on which B is constant. The right hand side is proportional to the current i enclosed by the circular loop. Ampere's circuital law is a generalisation of this result. We shall state this law for the restricted case when,

- (i) \mathbf{B} is directed along the tangent of the perimeter of a closed curve at every point;
- (ii) the magnitude of \mathbf{B} is constant along the perimeter. For this situation, the *Ampere's circuital law* states:

$$BL = \mu_0 i_e \quad (5.23)$$

where L is the perimeter of the closed curve or loop and i_e is the net current enclosed by the circuit. In other words, i_e is the net current crossing the surface defined by the perimeter of the closed circuit. The closed curve is called an *Amperean loop*. It is a geometrical entity and not a real wire loop.

Ampere's circuital law goes beyond the two restrictions mentioned above. We shall call these restrictions the *condition of tangency* and the *condition of constancy* (of B , the magnitude of \mathbf{B}). They apply only in situations of high symmetry. Although Ampere's circuital law holds for the case of the circular loop discussed in Section 5.3.2, it cannot be applied with profit to extract the simple expression $B = \mu_0 i / 2R$ [Eq. (5.21)] for the field at the centre of the loop. However, there exists a large number of situations of high symmetry where the theorem can be profitably applied. We shall see it in the next section to calculate the magnetic field produced by two commonly used and very useful magnetic systems: the *solenoid* and the *toroid*.

Ampere's circuital law is not distinct or different from Biot-Savart law. Both relate the magnetic field to the current, and both express the same physical consequences of electrical currents. Ampere's law is to Biot-Savart law, what Gauss' theorem is to Coulomb's law. Both, Ampere's and Gauss' law relate a physical quantity on the periphery (magnetic or electric field) to another physical quantity, namely, the source, in the interior (current or charge). We also note that Ampere's circuital law holds for steady currents which do not fluctuate with time. The example below will help us understand what is meant by the term "enclosed," current.

Example 5.4 Figure 5.9 shows a long straight wire of finite cross-section carrying steady current i . The current i is uniformly distributed across this cross-section of radius a . Calculate the magnetic field in the region $r < a$ and $r > a$.

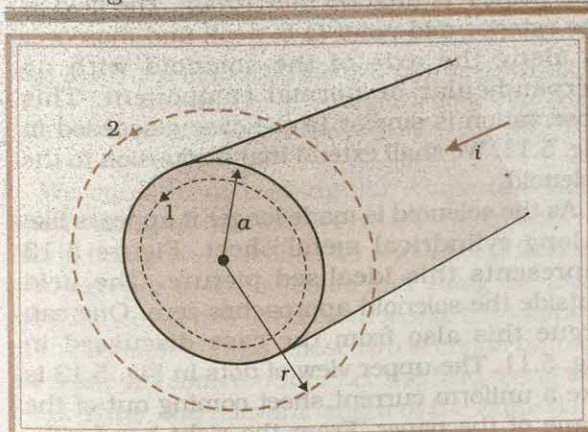


Fig. 5.9 A steady current i is uniformly distributed across a wire of radius a .

Answer Consider the case $r > a$. The Amperian loop is circle labelled 2. For this loop,

$$L = 2\pi r$$

$$i_e = i$$

in Eq. (5.23). The result is the familiar expression for a long straight wire

$$B(2\pi r) = \mu_0 i$$

$$B = \frac{\mu_0 i}{2\pi r} \quad (5.24)$$

$$B \propto \frac{1}{r} \quad (r > a)$$

Consider the case $r < a$. The Amperian loop is a circle labelled 1. For this loop, taking the radius of the circle to be r ,

$$L = 2\pi r$$

Now the current enclosed i_e is not i , but is less than this value. Since the current distribution is uniform, the fraction of i enclosed is,

$$i_e = i \left(\frac{\pi r^2}{\pi a^2} \right)$$

$$= \frac{ir^2}{a^2}$$

Thus from Eq. (5.23)

$$B(2\pi r) = \mu_0 \frac{ir^2}{a^2}$$

$$B = \left(\frac{\mu_0 i}{2\pi a^2} \right) r$$

$$B \propto r \quad (r < a)$$

Figure 5.10 shows a plot of the magnitude of \mathbf{B} with distance r from the centre of the wire. The

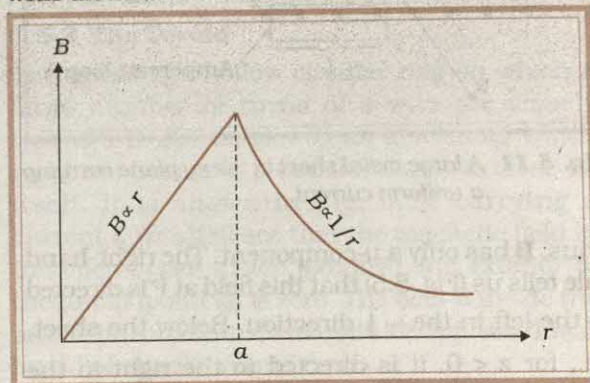


Fig. 5.10 Sketch of the magnitude of the magnetic field for the long conductor of radius a .

direction of the field is tangential to the circular loop 1 or 2 and given by the right-hand side rule described in Fig. 5.5. This example possesses the required symmetry so that Ampere's law can be applied readily.

5.5 THE SOLENOID AND THE TOROID

The solenoid and the toroid are two pieces of equipment which generate magnetic fields. In both, we come across a situation of high symmetry where Ampere's law can be profitably applied.

We begin our discussion by calculating the magnetic field due to a large metal sheet placed in the xy -plane and carrying a uniform surface current in the x -direction as shown in Fig. 5.11. The current in a width Δy is $K\Delta y$ and K is a constant. Note that K has dimensions of current per unit length. Let us consider a point P above the xy -plane. At this point, we argue that \mathbf{B} will have no x - or z -component ($B_x = B_z = 0$). From Biot-Savart law, we have that the x -component is zero since the current itself is in the x -direction. It also has no z -component since vertical component from a filament at $+y$ is cancelled by the corresponding filament at $-y$.

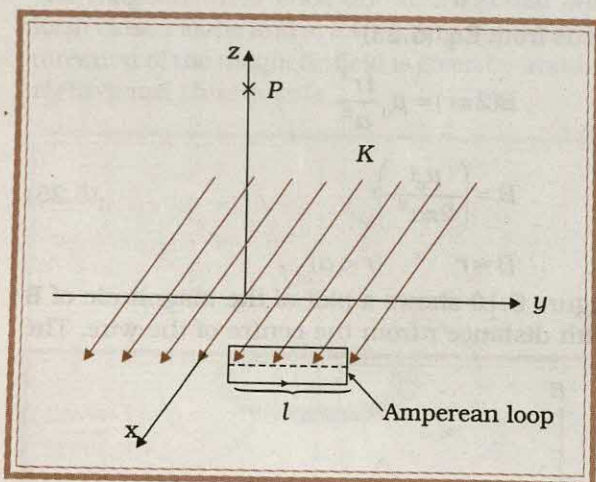


Fig. 5.11 A large metal sheet in the xy plane carrying a uniform current.

Thus, \mathbf{B} has only a y -component. The right-hand rule tells us (Fig. 5.5) that this field at P is directed to the left in the $-\hat{j}$ direction. Below the sheet, i.e., for $z < 0$, it is directed to the right in the $+\hat{j}$ direction. With these facts in mind, we draw a rectangular Amperian loop as shown in

Fig. 5.11. The vertical sections of this loop yield a null result since \mathbf{B} is in the $\pm \hat{j}$ direction while the sections are in the perpendicular $\pm \hat{k}$ directions. From Eq. (5.23) we have

$$L \doteq 2l$$

$$i_e = Kl$$

Thus,

$$B(2l) = \mu_0 Kl$$

$$B = \frac{\mu_0 K}{2} \quad (5.26)$$

Its direction is $+\hat{j}$ for $z < 0$ and $-\hat{j}$ for $z > 0$. The field is independent of the distance from the plane, just like the electric field of a uniform surface charge discussed in Section 1.15 of Chapter 1. The above discussion will aid us in understanding the field due to a solenoid.

5.5.1 The Solenoid

We shall discuss a long solenoid. By long solenoid we mean that the solenoid's length is large compared to its radius. It consists of a long wire wound in the form of a helix where the neighbouring turns are closely spaced. Each turn can be regarded as a circular loop and the net magnetic field is the vector sum of the fields due to all the turns.

Figure 5.12 displays the magnetic field lines for a finite solenoid. In Fig. 5.12(a), we show a section of this solenoid in an enlarged manner. In Fig. 5.12(b) we show the entire finite solenoid with its magnetic field. In Fig. 5.12(a), it is clear from the circular loops that the field lines tend to cancel between two neighbouring turns. In Fig. 5.12(b), we see that the field at the interior mid-point P is uniform and strong. The field at the exterior mid-point Q is weak and moreover is along the axis of the solenoid with no perpendicular or normal component. This observation is similar to the case discussed in Fig. 5.11. We shall extend that discussion to the solenoid.

As the solenoid is made longer it appears like a long cylindrical metal sheet. Figure 5.13 represents this idealised picture. The field outside the solenoid approaches zero. One can argue this also from the case discussed in Fig. 5.11. The upper view of dots in Fig. 5.13 is like a uniform current sheet coming out of the plane of the paper. From the right-hand rule, the field due to this is to the left at point Q (above) and to the right at point P (below). The lower row

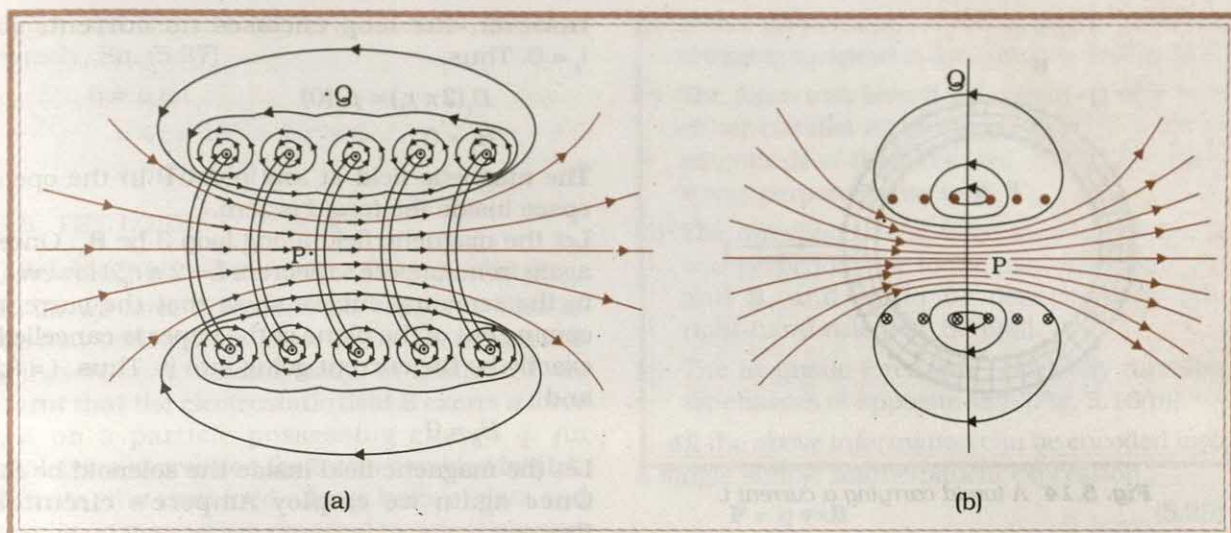


Fig. 5.12 (a) The magnetic field due to a section of the solenoid which has been stretched out for clarity. Only the interior semi-circular part is shown. Notice how the circular loops between neighbouring turns tend to cancel. (b) The magnetic field of a finite solenoid.

of crosses in Fig. 5.13 is like a uniform current sheet going into the plane of the paper. The field at any point above it is to the right. The two fields reinforce each other at P and cancel exactly at Q.

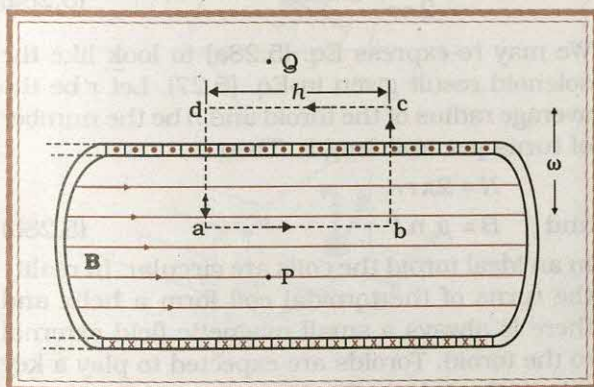


Fig. 5.13 The magnetic field of a very long solenoid. We consider a rectangular Amperian loop abcd to determine the field.

We consider a rectangular Amperian loop **abcd**. Along **cd** the field is zero as argued above. Along transverse sections **bc** and **ad**, the field is zero exterior to the solenoid and along the axis (and hence normal to the sections) in the interior of the solenoid. Thus, these two sections **bc** and **ad** make no contribution. Let the field along **ab** be **B**. Thus, the relevant length of the Amperian

loop is,

$$L = h$$

Let n be the number of turns per unit length. The enclosed current is,

$$i_e = i(nh)$$

where i is the current in the solenoid. From Ampere's circuital law [Eq. (5.23)]

$$BL = \mu_0 i_e$$

$$Bh = \mu_0 i(nh)$$

$$B = \mu_0 ni$$

The direction of the field is given by the right-hand rule. The solenoid is commonly used to obtain a uniform magnetic field. We shall see in the next chapter that a large field is possible by inserting a soft iron core inside the solenoid.

5.5.2 The Toroid

The toroid is a hollow circular ring on which a large number of turns of a wire are closely wound*. It can be viewed as a solenoid which has been bent into a circular shape to close on itself. It is shown in Fig. 5.14 carrying a current i . We shall see that the magnetic field in the open space inside (point P) and exterior to the toroid (point Q) is zero. The field **B** inside the toroid is constant in magnitude for the 'ideal' toroid of closely wound turns.

* The shape is similar to a doughnut or the Indian dish medu-wada.

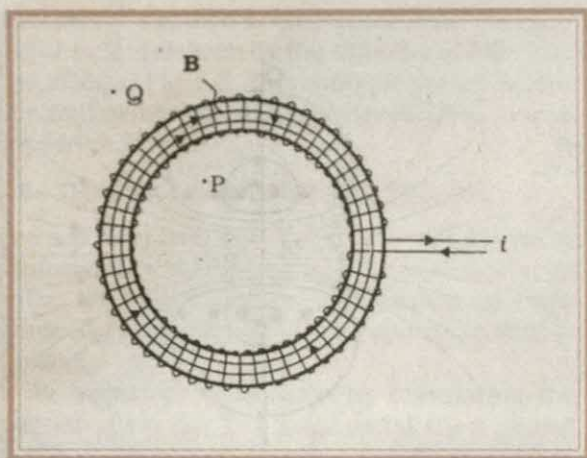


Fig. 5.14 A toroid carrying a current i .

Figure 5.15 shows a sectional view of the toroid. The direction of the magnetic field inside is clockwise as per the right-hand thumb rule for circular loops. Three circular Amperean loops are shown by dashed lines. By symmetry, the magnetic field should be tangential to them and constant in magnitude for each of the loops.

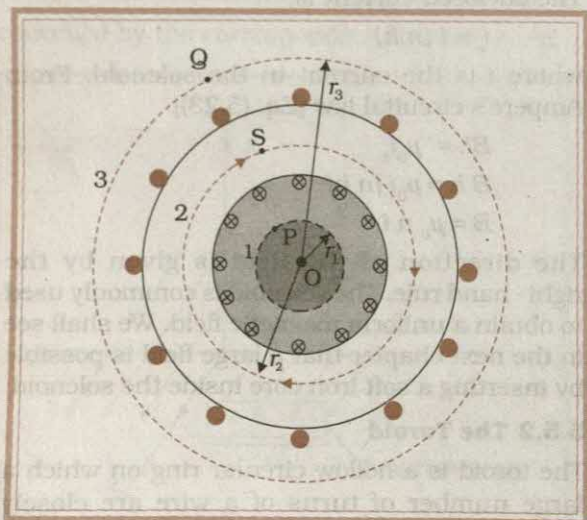


Fig. 5.15 A sectional view of the toroid. The magnetic field can be at an arbitrary distance r from the centre O of the toroid by Ampere's circuital theorem. The dashed lines labelled 1, 2 and 3 are three circular Amperean loops.

Let the magnetic field along loop 1 be B_1 in magnitude. Then from Ampere's circuital law (Eq. 5.23)

$$L = 2\pi r_1$$

However, the loop encloses no current, so $i_e = 0$. Thus,

$$B_1(2\pi r_1) = \mu_0(0)$$

$$B_1 = 0$$

The magnetic field at any point P in the open space inside the toroid is zero.

Let the magnetic field along loop 3 be B_3 . Once again from Ampere's theorem $L = 2\pi r_3$. However, in the sectional cut we show that the current coming out of the plane of the paper is cancelled exactly by the current going into it. Thus, $i_e = 0$, and

$$B_3 = 0$$

Let the magnetic field inside the solenoid be B . Once again we employ Ampere's circuital theorem,

$$L = 2\pi r$$

The current enclosed i_e is given by N turns of toroidal coils and is Ni . Thus,

$$i_e = Ni$$

$$B(2\pi r) = \mu_0 Ni$$

$$B = \frac{\mu_0 Ni}{2\pi r} \quad (5.28a)$$

We may re-express Eq. (5.28a) to look like the solenoid result given in Eq. (5.27). Let r be the average radius of the toroid and n be the number of turns per unit length. Then

$$N = 2\pi r n$$

$$\text{and } B = \mu_0 n i \quad (5.28b)$$

In an ideal toroid the coils are circular. In reality the turns of the toroidal coil form a helix and there is always a small magnetic field external to the toroid. Toroids are expected to play a key role in the **tokamak**, an equipment for plasma confinement in fusion power reactors.

Example 5.5 A solenoid of length 0.5 m has a radius of 1 cm and is made up of 500 turns. It carries a current of 5 A. What is the magnitude of the magnetic field inside the solenoid?

Answer The number of turns per unit length is,

$$n = \frac{500}{0.5} = 1000 \text{ turns/m}$$

The length $l = 0.5$ m and radius $r = 0.01$ m. Thus, $l/a = 50$ i.e., $l \gg a$.

Hence, we can use the *long solenoid* formula, namely, Eq. (5.27)

$$\begin{aligned} B &= \mu_0 n i \\ &= 4\pi \times 10^{-7} \times 10^3 \times 5 \\ &= 6.28 \times 10^{-3} \text{ T} \end{aligned}$$

5.6 THE LORENTZ FORCE

5.6.1 Magnetic Force on a Moving Charge

So far we have discussed the magnetic field \mathbf{B} . We now address the issue of the force which \mathbf{B} will exert on a material body. In Chapter 1, we learnt that the electrostatic field \mathbf{E} exerts a force $q\mathbf{E}$ on a particle possessing charge q . An analogous expression for \mathbf{B} was provided by H.A. Lorentz who examined a large body of data and results on magnetism reported by scientists ever since Oersted's discovery. Some of the important features of force \mathbf{F} were:

- F was proportional to the magnitudes of the charge q , its speed v , and the magnetic field B .
- The force was zero if the direction of \mathbf{v} was either parallel or anti-parallel to \mathbf{B} and the magnitude of the force was maximum when \mathbf{v} was perpendicular to \mathbf{B} .
- The direction of the force was *sideways*. It was perpendicular to the plane defined by \mathbf{v} and \mathbf{B} , and could be described by the right-hand rule [Fig. 5.16(a)].
- The magnetic force was oppositely directed for charges of opposite sign [Fig. 5.16(b)].

All the above information can be encoded into a single simple mathematical expression,

$$\mathbf{F} = q \mathbf{v} \times \mathbf{B} \quad (5.29)$$

Using the properties of cross-product, we can verify that Eq. (5.29) is consistent with the observations on magnetic force made above. The magnitude of the force is

$$F = qvB \sin \theta$$

Thus, F is zero if $\theta = 0$ (\mathbf{v} parallel to \mathbf{B}) or $\theta = \pi$ (\mathbf{v} anti-parallel to \mathbf{B}). It is maximum when $\theta = \pi/2$.

We shall see that Eq. (5.29) implies that the magnetic force acting on a charged particle does no work. It cannot change the speed v of the particle. To see this, note that the instantaneous power is zero,

$$\begin{aligned} \mathbf{F} \cdot \mathbf{v} &= q \mathbf{v} \cdot (\mathbf{v} \times \mathbf{B}) \\ &= 0 \end{aligned}$$

From Newton's second Law, the left hand side is,

$$m \frac{d\mathbf{v}}{dt} \cdot \mathbf{v} = 0$$

We could write this as

$$\frac{m}{2} \frac{d}{dt} (\mathbf{v} \cdot \mathbf{v}) = 0$$

where we have used the chain rule. Thus

$$\frac{d}{dt} \left(\frac{m}{2} v^2 \right) = 0$$

However, the kinetic energy is $K = mv^2/2$. So,

$$\frac{dK}{dt} = 0$$

The kinetic energy of the charged particle does not change. This means that the *speed* of the particle remains constant. Recall also, the

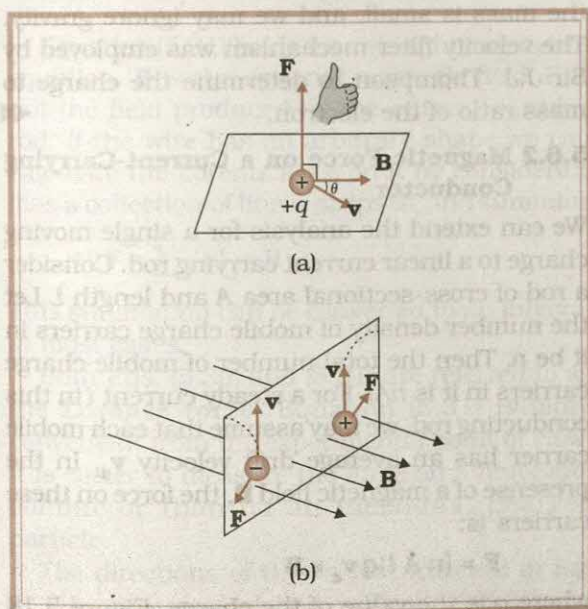


Fig. 5.16 The direction of the magnetic force acting on a charged particle. (a) The force on a positively charged particle with velocity \mathbf{v} and making an angle θ with the magnetic field \mathbf{B} is given by the right-hand rule. (b) A moving charged particle q is deflected in an opposite sense to $-q$ in the presence of magnetic field.

- F was zero on a charge q if the charge was at rest. Just as moving charges produce a magnetic field (Biot-Savart), so too only moving charges are affected by a magnetic field.

work-energy theorem: the change in kinetic energy is equal to the work done by the net force on the particle. In our case there is only a single force acting on the charged particle, namely, the magnetic force. Thus, the *magnetic force does no work*.

As mentioned earlier, the unit of the magnetic field is tesla (T). From Eq. (5.29) its dimension is readily determined to be $M^1T^{-2}A^{-1}$, where A represents current. Many textbooks use Eq. (5.29) instead of the Biot-Savart law [Eq. (5.41)] to define the magnetic field.

In the presence of an electric field \mathbf{E} and magnetic field \mathbf{B} , the total force on a moving charged particle is

$$\mathbf{F} = q(\mathbf{v} \times \mathbf{B} + \mathbf{E}) \quad (5.30)$$

This expression is called the *Lorentz force*, and is named after the Dutch physicist Hendrik Anton Lorentz.

Example 5.6 The Velocity Filter: A stream of charged particles possessing a range of speeds enters region I after passing through a slit S_1 (Fig. 5.17). In region I there exists crossed (perpendicular) electric and magnetic fields. The electric field has magnitude 100 V/m. We want the particles emerging from slit S_2 into region II to have a fixed velocity of 1000 m/s. What should be the value of the uniform magnetic field in region I?

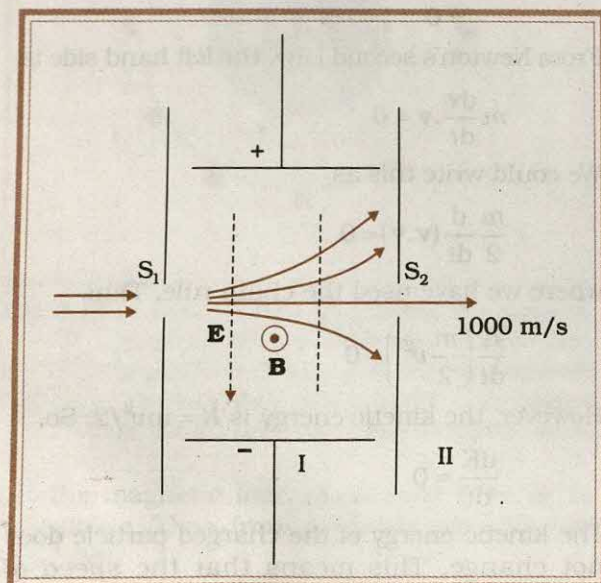


Fig. 5.17 Figure for Example 5.6.

Answer The magnetic field \mathbf{B} will impart a sideways force to the charged particles entering region I. For particles to go undeflected the crossed electric and magnetic fields must balance each other. Thus, from Eq. (5.30) for the Lorentz force,

$$qE = qvB$$

$$\text{or} \quad v = \frac{E}{B}$$

It is given that $E = 100$ V/m and it is required that $v = 1000$ m/s. Thus,

$$B = \frac{E}{v} \\ = 0.1 \text{ T}$$

Note that the magnitude of the charge is of no consequence to the final answer. For elementary charged particles such as the electron or proton, the mass is small, and we may ignore gravity. The velocity filter mechanism was employed by Sir J.J. Thompson to determine the charge to mass ratio of the electron. ◀

5.6.2 Magnetic Force on a Current-Carrying Conductor

We can extend the analysis for a single moving charge to a linear current carrying rod. Consider a rod of cross-sectional area A and length l . Let the number density of mobile charge carriers in it be n . Then the total number of mobile charge carriers in it is nAl . For a steady current i in this conducting rod, we may assume that each mobile carrier has an average drift velocity \mathbf{v}_d . In the presence of a magnetic field \mathbf{B} , the force on these carriers is:

$$\mathbf{F} = (nAl)q\mathbf{v}_d \times \mathbf{B}$$

where q is the value of the charge. Figure 5.18 illustrates this situation. Now $nq\mathbf{v}_d$ is the current density \mathbf{j} and $(nq\mathbf{v}_d)A$ is the current i (Chapter 3 for the discussion of current and current density). Thus

$$\begin{aligned} \mathbf{F} &= (nq\mathbf{v}_d)Al \times \mathbf{B} \\ &= \mathbf{j}Al \times \mathbf{B} \\ &= (\mathbf{j}A)l \times \mathbf{B} \\ &= i\mathbf{l} \times \mathbf{B} \end{aligned} \quad (5.31)$$

where \mathbf{l} is a vector of magnitude l , the length of the rod, and direction the same as the current i . Note that the current i is not a vector. In the last

step leading to Eq. (5.31), we have transferred the vector sign from \mathbf{j} to \mathbf{l} .

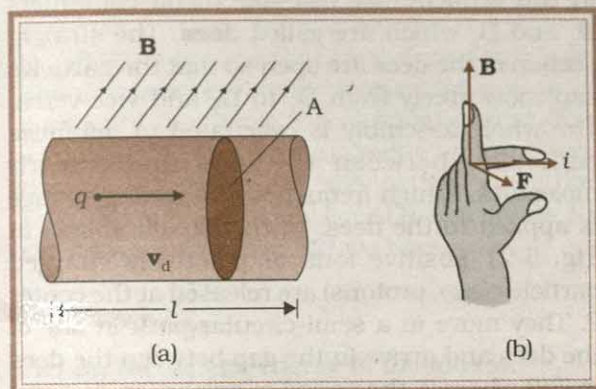


Fig. 5.18 (a) A section of a straight conducting rod carrying current i and placed in an external magnetic field \mathbf{B} . (b) Fleming's left-hand rule.

Equation (5.31) holds for a straight rod. In this equation, \mathbf{B} is the external magnetic field. It is not the field produced by the current carrying rod. If the wire has an arbitrary shape we can calculate the Lorentz force on it by considering it as a collection of linear strips $d\mathbf{l}$, and summing

$$\mathbf{F} = \sum_j i d\mathbf{l}_j \times \mathbf{B}$$

This summation can be converted to an integral in most cases.

Both Eqs. (5.30) and (5.31) are referred to as the **Lorentz force**. Equation (5.31) is more convenient from the experimental point of view. It is easier to measure the force on a finite size conductor than on an individual charged particle.

The directions of the vectors involved in Eq. (5.31) can be remembered by Fleming's left-hand rule. Extend the middle finger, forefinger and the thumb of your left hand in mutually perpendicular directions. Then, if the middle finger is along the current, the forefinger is along the field, the thumb gives the direction of the force. The rule is illustrated in Fig. 5.18(b).

Example 5.7 A straight wire of mass 200 g and length 1.5 m carries a current of 2 A. It is suspended in mid-air by a uniform horizontal magnetic field \mathbf{B} . What is the magnitude of the magnetic field?

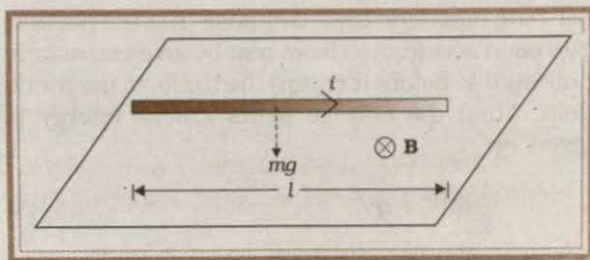


Fig. 5.19 Illustration of Example 5.7.

Answer From Eq. (5.31) we have that for mid-air suspension

$$mg = ilB$$

$$B = \frac{mg}{il}$$

$$= \frac{0.2 \times 9.81}{2 \times 1.5} = 0.65 \text{ T}$$

Note that it would have been sufficient to specify m/l , the mass per unit length of the wire. The earth's magnetic field is approximately 4×10^{-5} T and we have ignored it. \leftarrow

5.7 THE CYCLOTRON

The cyclotron is a machine to accelerate charged particles or ions to high energies. It was invented by E.O. Lawrence and M.S. Livingston in 1934 to investigate nuclear structure. The cyclotron uses both electric and magnetic fields to generate energetic particles.

To understand how a cyclotron works, we must first understand the motion of a charged particle in a perpendicular uniform magnetic field. Figure 5.20 illustrates this for a positively charged particle. The particle moves in a circle with its plane perpendicular to the field \mathbf{B} . This is clear from Eq. (5.29). The force is invariably *sideways* to the motion of the particle. We have also learnt in the previous section that this force will not change the speed of the particle. Hence, the particle performs uniform circular motion. From Newton's second law and Eq. (5.29)

$$qvB = \frac{mv^2}{r}$$

where v^2/r is the centripetal acceleration of the particle. Thus,

$$r = \frac{mv}{qB} \quad (5.32)$$

The radius of the circular orbit scales inversely with the charge to mass ratio (q/m) and inversely

with the magnetic field. Suppose that the particle has been accelerated from rest by an electrostatic potential V before it enters the uniform magnetic field. Then the change in its kinetic energy is given by

$$\frac{1}{2}mv^2 = qV \quad (5.33)$$

where we have used the work-energy theorem. Employing Eq. (5.33) in Eq. (5.32) we have

$$r = \frac{1}{B} \left(\frac{2mV}{q} \right)^{1/2} \quad (5.34)$$

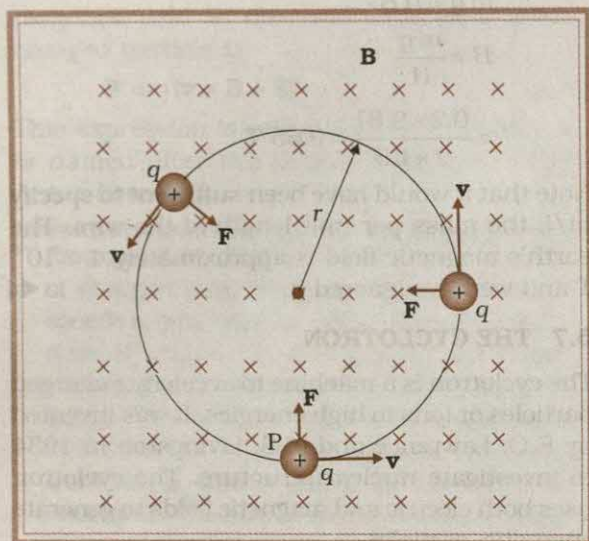


Fig. 5.20 A positively charged particle moving in a uniform magnetic field directed into the plane of the paper. The Lorentz force \mathbf{F} is invariably directed towards the center of the circle.

Since the particle performs uniform circular motion, we may express $v = \omega_c r$ where ω_c is the angular frequency. The subscript c stands for cyclotron. Employing $v = \omega_c r$ in Eq. (5.32),

$$\begin{aligned} r &= \frac{m\omega_c r}{qB} \\ \omega_c &= \frac{qB}{m} \\ f_c &= \frac{qB}{2\pi m} \end{aligned} \quad (5.35)$$

where f_c is called the *cyclotron frequency*. The point of interest is that the *cyclotron frequency does not depend on the speed of the particle*. This fact is exploited in the design of the cyclotron.

Figure 5.21 shows a schematic view of the cyclotron. The motion of charged particles occurs in two semicircular disc-like metal containers D_1 and D_2 which are called **dees**. The straight section of the dees are open so that the particles can move freely from D_1 to D_2 and vice-versa. The whole assembly is evacuated to minimise collisions between the ions and the air molecules. A high frequency alternating voltage is applied to the dees. In the sketch shown in Fig. 5.21 positive ions or positively charged particles (e.g., protons) are released at the center P . They move in a semi-circular path in one of the dees and arrive in the gap between the dees in $T/2$ where T , the period of revolution, is given by Eq. (5.35),

$$T = \frac{1}{f_c} = \frac{2\pi m}{qB} \quad (5.36)$$

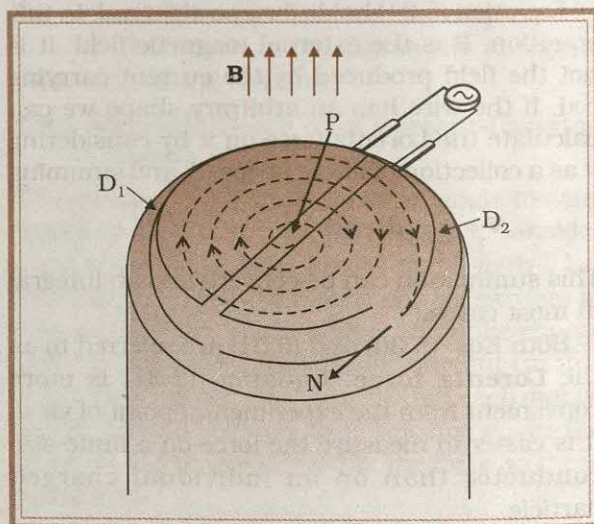


Fig. 5.21 A schematic sketch of the cyclotron. There is a source of charged particles or ions at P which move in a circular fashion in the dees, D_1 and D_2 , on account of a uniform perpendicular magnetic field B . An alternating voltage source accelerates these ions to high speeds. The ions are eventually 'extracted' at the point N .

The frequency f_a of the applied voltage is adjusted so that the polarity of the dees is reversed in the same time that it takes the ions to complete one half of the revolution. The requirement $f_a = f_c$ is called the *resonance condition*. The phase of the supply is adjusted

so that when the positive ions arrive at the edge of D_1 , D_2 is at a lower potential ($V_{D_1} - V_{D_2} = V$) and the ions are accelerated across the gap. The increase in their kinetic energy is qV . From Eq. (5.32) it is clear that their radius increases. The ions are repeatedly accelerated across the dees until they have the required energy. They are then deflected by a magnetic field and leave the system via an exit slit. At this stage their radius is approximately equal to the radius R of the dees. From Eq. (5.32) we have,

$$v = \frac{qBR}{m}$$

Hence, the kinetic energy of the ions is,

$$\frac{1}{2}mv^2 = \frac{q^2 B^2 R^2}{2m} \quad (5.37)$$

The operation of the cyclotron is based on the fact that the time for one revolution of the ions is independent of the speed or radius [Eq. (5.35) or (5.36)]. The cyclotron was used to bombard nuclei with energetic particles and study the resulting nuclear reactions. It is used to implant ions into solids and modify their properties or even synthesise new materials. It is used in hospitals to produce radioactive substances which can be used in diagnosis and treatment.

The cyclotron does have limitations. At high speeds relativistic effects must be accounted for. The mass of the ion is no longer constant and the delicate resonance condition, $f_a = f_c = qB/(2\pi m)$ is upset. Light elementary particles such as electrons require unusually high frequencies (GHz). Maintaining the uniformity of the magnetic field over the extended region of the dees can also be a problem. It is for these reasons that other accelerating machines such as the synchrotron have been developed.

5.8 THE AMPERE

We have learnt that there exists a magnetic field due to a conductor carrying a current. This follows from the Biot-Savart law or Ampere's circuital law. Further, we have learnt that an external magnetic field will exert a force on a current carrying conductor. This follows from the Lorentz force formula. Thus, it is logical to expect that two current carrying straight conductors placed near each other will exert (magnetic) forces on each other. In the period 1820-25,

Ampere studied the nature of this magnetic force and its dependence on the magnitude of the current, the shape and size of the conductors as well as the distances between the conductors. In this section we shall take the simple example of two parallel current bearing conductors, which will perhaps help us to appreciate Ampere's painstaking work.

Figure 5.22 shows two long parallel conductors separated by a distance d and bearing currents i_a and i_b . The conductor 'a' produces a magnetic field \mathbf{B}_a at all points along the conductor 'b'. The right-hand rule (Fig. 5.5) tells us that the direction of this field is downwards. Its magnitude is given by Eq. (5.16) or from Ampere's circuital law,

$$B_a = \frac{\mu_0 i_a}{2\pi d}$$

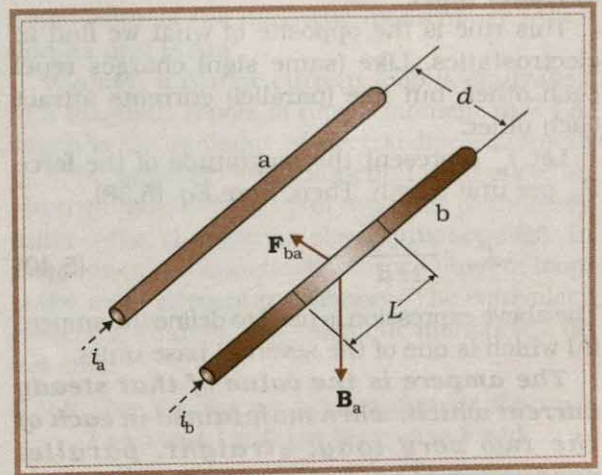


Fig. 5.22 Two long straight parallel conductors carrying steady currents i_a and i_b and separated by a distance d . \mathbf{B}_a is the magnetic field set up by conductor 'a' at conductor 'b'.

The conductor 'b' will experience a sideways force on account of the external field \mathbf{B}_a . The direction of this force is towards the conductor 'a'. You can verify this either by the cross product rule of vectors or by Fleming's left hand rule which is depicted in Fig. 5.18(b). We label this force \mathbf{F}_{ba} , the force on a segment L of 'b' due to 'a'. The magnitude of this force is given by Eq. (5.31),

$$\begin{aligned} F_{ba} &= i_b L B_a \\ &= \frac{\mu_0 i_a i_b}{2\pi d} L \end{aligned} \quad (5.38)$$

It is of course possible to compute the force on 'a' due to 'b'. From considerations similar to above we find that the force \mathbf{F}_{ab} , on a segment of length L of 'a' due to the current in 'b' is equal in magnitude to \mathbf{F}_{ba} , and directed towards 'b'. Thus,

$$\mathbf{F}_{ba} = -\mathbf{F}_{ab} \quad (5.39)$$

Note that this is consistent with Newton's third Law. Thus, at least for parallel conductors and steady currents, we have shown that the Biot-Savart law and the Lorentz force yield results in accordance with Newton's third Law.

We have seen from above that currents flowing in the same direction attract each other. One can show that oppositely directed currents repel each other. Thus,

Parallel currents attract, and antiparallel currents repel.

This rule is the opposite of what we find in electrostatics. Like (same sign) charges repel each other, but like (parallel) currents attract each other.

Let f_{ba} represent the magnitude of the force \mathbf{F}_{ba} per unit length. Then, from Eq. (5.38),

$$f_{ba} = \frac{\mu_0 i_a i_b}{2\pi d} \quad (5.40)$$

The above expression is used to define the ampere (A) which is one of the seven SI base units.

The ampere is the value of that steady current which, when maintained in each of the two very long, straight, parallel conductors of negligible cross-section, and placed one meter apart in vacuum, would produce on each of these conductors a force equal to 2×10^{-7} newtons per metre of length.

This definition of the ampere was adopted in 1946. It is a *theoretical* definition. In practice one must eliminate the effect of the earth's magnetic field and substitute *very long* wires by multiturn coils of appropriate geometries. An instrument called the *current balance* is used to measure this mechanical force.

The SI unit of charge, namely, the *Coulomb*, can now be defined in terms of the ampere.

When a steady current of 1A is set up in a conductor, the quantity of charge that flows through its cross-section in 1s is one coulomb (1C).

Example 5.8 The horizontal component of the earth's magnetic field at a certain place is 3.0×10^{-5} T and the direction of the field is from the geographic south to the geographic north. A very long straight conductor is carrying a steady current of 1A. What is the force per unit length on it when it is placed on a horizontal table and the direction of the current is (a) east to west; (b) south to north?

Answer $\mathbf{F} = i\mathbf{l} \times \mathbf{B}$ [Eq. (5.31)]

$$F = ilB \sin \theta$$

The force per unit length is

$$f = F/l = iB \sin \theta$$

- (a) When the current is flowing from east to west, $\theta = 90^\circ$

Hence,

$$\begin{aligned} f &= iB \\ &= 1 \times 3 \times 10^{-5} = 3 \times 10^{-5} \text{ N m}^{-1} \end{aligned}$$


This is larger than the value $2 \times 10^{-7} \text{ Nm}^{-1}$ quoted in the definition of the ampere. Hence it is important to eliminate the effect of the earth's magnetic field and other stray fields while standardising the ampere.

The direction of the force is downwards. This direction may be obtained either by Fleming's left hand rule [Fig. 5.18(b)] or the directional property of cross product of vectors.

- (b) When the current is flowing from south to north,

$$\theta = 0^\circ$$

$$F = 0$$

Hence no force per unit length on the conductor. 

5.9 THE CURRENT LOOP AS A MAGNETIC DIPOLE

The magnetic field in the neighbourhood of a finite solenoid bears a close resemblance to that of a bar magnet. One may see this by moving a compass needle around these two bodies and noting that the deflections of the needle are similar in both cases. Ampere was one of the first to make such comparisons.

He took a simpler current element, the current loop, and demonstrated its resemblance to a bar magnet. He put forward the bold hypothesis that all magnetic phenomena is due to circulating electrical currents. It turns out that Ampere's hypothesis is basically correct.

In this section we consider the elementary magnetic element: the current loop. We show that the magnetic field due to current in a circular conductor has a dipole character. We next demonstrate that a rectangular current carrying wire placed in a uniform magnetic field experiences a torque, but no net force. This situation should remind us of an electric dipole in a uniform electric field of section 1.10, Chapter 1. Finally, we show that an electron revolving around a positively charged nucleus resembles an elementary magnetic dipole and we calculate its dipole moment.

5.9.1 The Circular Current Loop as a Magnetic Dipole

In section 5.3.2, we have evaluated the magnetic field on the axis of a circular loop carrying a steady current i . The magnitude of this field was [Eq. (5.20)],

$$B = \frac{\mu_0 i R^2}{2(x^2 + R^2)^{3/2}}$$

and its direction was along the axis and given by the right-hand thumb rule (Fig. 5.8). We remind ourselves that R is the radius of the loop and x the distance along the axis as measured from the center of the loop. For $x \gg R$, we may drop the R^2 term in the denominator. Thus,

$$B \approx \frac{\mu_0 i R^2}{2x^3}$$

Note that the area of the loop $A = \pi R^2$. Thus,

$$B \approx \frac{\mu_0 i A}{2\pi x^3}$$

We define the magnetic moment \mathbf{m} to have a magnitude iA , i.e., the product of the current and its loop area. Its direction is defined to be normal to the plane of the loop in the sense given by the right hand thumb rule. Further, we consider the area \mathbf{A} to be a vector whose direction is perpendicular to the planar loop and given by the same right-hand thumb rule, viz. Fig. 5.8, which indicates the direction of the magnetic field. So,

$$\mathbf{m} = i \mathbf{A} \quad (5.41)$$

$$\begin{aligned} \mathbf{B} &\approx \frac{\mu_0 \mathbf{m}}{2\pi x^3} \\ &= \frac{\mu_0}{4\pi} \frac{2\mathbf{m}}{x^3} \end{aligned} \quad (5.42)$$

Now, let us make the substitution

$$\mu_0 \rightarrow 1/\epsilon_0$$

$$\mathbf{m} \rightarrow \mathbf{p}_e \quad (\text{electrostatic dipole})$$

$$\mathbf{B} \rightarrow \mathbf{E} \quad (\text{electrostatic field})$$

we obtain,

$$\mathbf{E} = \frac{2\mathbf{p}_e}{4\pi\epsilon_0 x^3}$$

which is precisely the axial field for an electric dipole considered in Chapter 1, Section 1.9 [Eq. (1.25)].

The equatorial field analogy is not described since the evaluation of such a field for a current loop turns out to be very difficult. We can, however, derive the axial result Eq. (5.42) for a square loop. This *axial result holds for any planar current loop*. As is clear from Eq. (5.41), the dimensions of the magnetic moment are $[A][L^2]$ and its unit is Am^2 .

In short, a planar current loop is equivalent to a magnetic dipole of dipole moment $\mathbf{m} = i \mathbf{A}$, which is the analogue of electric dipole moment \mathbf{p} . Note, however, a fundamental difference: an electric dipole is built up of two more elementary units—the charges (or electric monopoles). In magnetism, a magnetic dipole (or a current loop) is the most elementary element. The equivalent of electric charges, i.e., magnetic monopoles, do not exist.

5.9.2 The Torque on a Rectangular Current Loop in a Uniform Magnetic Field

We now show that a rectangular loop carrying a steady current i and placed in a uniform magnetic field experiences a torque. It does not experience a net force. This behaviour is analogous to that of electric dipole in a uniform electric field (Section 1.10).

We first consider the simple case when the rectangular loop is placed such that the uniform magnetic field \mathbf{B} is in the plane of the loop. This is illustrated in Fig. 5.23(a).

The field exerts no force on the two arms AB and CD of the loop. It is perpendicular to the arm BC of the loop and exerts a force \mathbf{F}_2 on it which is directed into the plane of the loop. Its magnitude is,

$$F_2 = i b B$$

Similarly it exerts a force \mathbf{F}_1 on the arm DA and \mathbf{F}_1 is directed out of the plane of the paper.

$$F_1 = i b B = F_2$$

Thus, the *net force* on the loop is zero. There is a torque on the loop due to the pair of forces \mathbf{F}_1 and \mathbf{F}_2 . fig. 5.23(b) shows a view of the loop from the bottom end. It shows that the torque on the loop tends to rotate it clockwise. This torque is (in magnitude),

$$\begin{aligned}\tau &= F_1 \frac{a}{2} + F_2 \frac{a}{2} \\ &= i b B \frac{a}{2} + i b B \frac{a}{2} = i(ab)B \\ &= iAB\end{aligned}\quad (5.43)$$

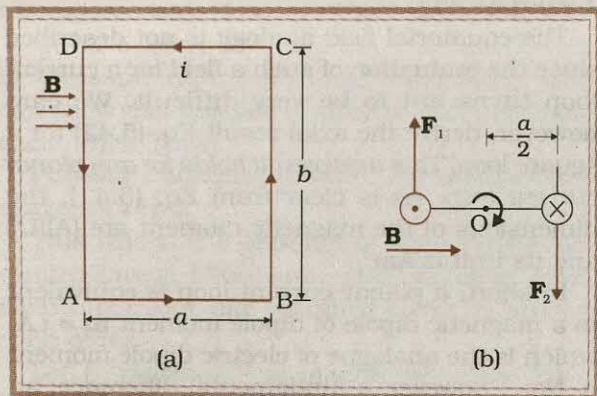


Fig. 5.23 (a) The uniform magnetic field is in the plane of the rectangular loop and parallel to the sides AB and CD. (b) Sectional view of the loop from the lower end.

where $A = ab$ is the area of the rectangle. We define the magnetic moment of the current loop in the same way as in Eq. (5.41),

$$\mathbf{m} = i\mathbf{A}$$

where the direction of the area vector \mathbf{A} is given by the right-hand thumb rule and is directed out of the plane of the paper in Fig. 5.23. Then, from Eq. (5.43),

$$\tau = \mathbf{m} \times \mathbf{B} \quad (5.44)$$

which is analogous to the electrostatic case,

$$\tau = \mathbf{p}_e \times \mathbf{E}$$

Equation (5.44) can be proved for the general case when the angle θ between the magnetic field \mathbf{B} and the area vector \mathbf{A} is arbitrary and not 90° as in Fig. 5.23. Figure 5.24 illustrates this general case. The forces on the arms DC and BA are equal, opposite, and collinear. They cancel resulting in no net force or torque. The forces on arms AD and CB are \mathbf{F}_1 and \mathbf{F}_2 . They too are equal and opposite, with magnitude,

$$F_1 = F_2 = i b B$$

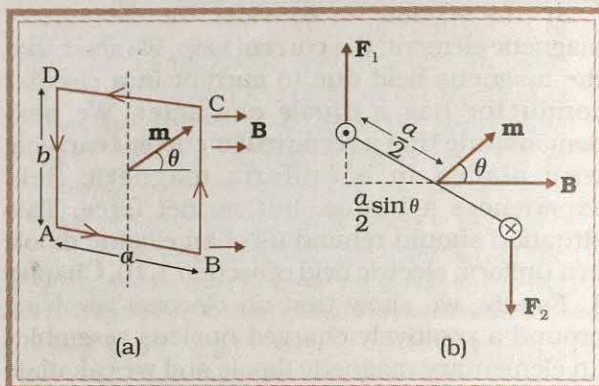


Fig. 5.24 (a) The area vector of the loop ABCD makes an arbitrary angle θ with the magnetic field. (b) Top view of the loop. The forces \mathbf{F}_1 and \mathbf{F}_2 acting on the arms AD and CB are indicated. The arms AD and CB are shown by the symbols \odot , and \otimes respectively.

But they are not collinear! This results in a couple. Figure 5.24(b) is a top view of the arrangement and it illustrates these two forces constituting a couple. The magnitude of the torque on the loop is,

$$\begin{aligned}\tau &= F_1 \frac{a}{2} \sin \theta + F_2 \frac{a}{2} \sin \theta \\ &= IabB \sin \theta \\ &= IAB \sin \theta \\ &= |\mathbf{m}|B \sin \theta\end{aligned}$$

The torque tends to rotate the loop clockwise about the dashed line shown in Fig. 5.24(a). We see that

$$\tau = \mathbf{m} \times \mathbf{B} \quad (5.44)$$

once again. If the loop has N closely wound turns, Eq. (5.44) still holds, with

$$\mathbf{m} = N i \mathbf{A} \quad (5.45)$$

The current loop tends to rotate in an external uniform magnetic field. The dynamical formula for this is codified in Eq. (5.44). The result is significant. We note that a compass needle also tends to rotate in an external field. Once again we see the similar behaviour of a current loop and a magnetic needle. This property of rotation is exploited in the electric motor, a device we encounter over a dozen times in a single day. Think of Eq. (5.44) the next time you switch on a fan.

5.9.3 The Magnetic Dipole Moment of a Revolving Electron

In Chapter 13 we shall read about the Bohr model of the hydrogen atom. You may perhaps have heard of this model which was proposed by the Danish physicist Niels Bohr in 1911 and was a stepping stone to a new kind of mechanics, namely, quantum mechanics. In the Bohr model, the electron is a negatively charged particle revolving around a positively charged nucleus much as a planet revolves around the sun. The force in the former case is electrostatic (Coulomb force) while it is gravitational for the planet - Sun case. We show this Bohr picture of the electron in Fig. 5.25.

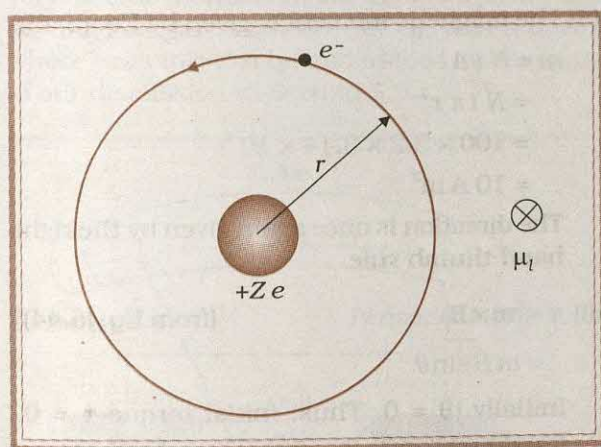


Fig. 5.25 In the Bohr model of hydrogen-like atoms, the negatively charged electron is revolving with uniform speed around a centrally placed positively charged $(+Ze)$ nucleus. The uniform circular motion of the electron constitutes a current. The direction of the magnetic moment is into the plane of the paper and is indicated separately by \otimes .

The electron of charge $-e$ ($e = +1.6 \times 10^{-19}$ C) performs uniform circular motion around a stationary heavy nucleus of charge $+Ze$. This constitutes a current i , where,

$$i = \frac{e}{T} \quad (5.46)$$

and T is the time period of revolution. Let r be the orbital radius of the electron, and v the orbital speed. Then,

$$T = \frac{2\pi r}{v} \quad (5.47)$$

Substituting Eq. (5.47) in Eq. (5.46), we have

$$i = \frac{ev}{2\pi r}$$

There will be a magnetic moment μ_l associated with this circulating current. From Eq. (5.41) its magnitude is,

$$\begin{aligned} \mu_l &= i \pi r^2 \\ &= \frac{evr}{2} \end{aligned}$$

The direction of this magnetic moment is into the plane of the paper in Fig. 5.25. This follows from the right hand rule discussed earlier and the fact that the negatively charged electron is shown as moving anti-clockwise in Fig. 5.25. The current is therefore clockwise. Multiplying and dividing the right hand side of the above expression by the electron mass m_e , we have,

$$\begin{aligned} \mu_l &= \frac{e}{2m_e} (m_e v r) \\ &= \frac{e}{2m_e} l \end{aligned} \quad (5.48)$$

Here, l is the magnitude of the angular momentum of the electron about the central nucleus. Vectorially,

$$\mu_l = -\frac{e}{2m_e} l$$

The negative sign indicates that the angular momentum of the electron is opposite in direction to the magnetic moment. The ratio

$$\frac{\mu_l}{l} = \frac{e}{2m_e} \quad (5.49)$$

is called the *gyromagnetic ratio* and is a constant. Its value is 8.8×10^{10} C/kg.

The fact that even at an atomic level there is a magnetic moment confirms Ampere's bold hypothesis of atomic magnetic moments. This according to Ampere would help one to explain the magnetic properties of materials. Can one assign a value to this atomic dipole moment? The answer is one can do so within the Bohr model. Bohr hypothesised that the angular momentum assumes a discrete set of values, namely,

$$l = \frac{nh}{2\pi} \quad (5.50)$$

where n is a natural number, $n = 1, 2, 3, \dots$ and h is a constant named after Max Planck (Planck's constant) with a value $h = 6.626 \times 10^{-34}$ J s.

This condition of discreteness is called the *Bohr quantisation condition*. We shall discuss it in detail in Chapter 13. Our aim here is merely to use it to calculate the elementary dipole moment. Take the value $n = 1$, we have from Eq. (5.48) that,

$$\begin{aligned}(\mu_l)_{\min} &= \frac{e}{4\pi m_e} h \\&= \frac{1.60 \times 6.63 \times 10^{-19} \times 10^{-34}}{4 \times 3.14 \times 9.11 \times 10^{-31}} \\&= 9.27 \times 10^{-24} \text{ A m}^2\end{aligned}\quad (5.51)$$

where the subscript 'min' stands for minimum. This value is called the *Bohr magneton*.

Any charge in uniform circular motion would have an associated magnetic moment given by an expression similar to Eq. (5.48). This dipole moment is labelled as the *orbital magnetic moment*. Hence the subscript 'l' in μ_l . Besides the orbital moment, the electron has an *intrinsic magnetic moment*, which has the same numerical value as given in Eq. (5.51). It is called the *spin magnetic moment*. But we hasten to add that it is not as though the electron is spinning. The electron is an elementary particle and it does not have an axis to spin around like a top or our earth. Nevertheless it does possess this *intrinsic magnetic moment*. The microscopic roots of magnetism in iron and other materials can be traced back to this intrinsic spin magnetic moment.

Example 5.9 A 100 turn closely wound circular coil of radius 10 cm carries a current of 3.2 A. (i) What is the field at the center of the coil? (ii) What is the magnetic moment of this arrangement?

The coil is placed in a vertical plane and is free to rotate about a horizontal axis which coincides with its diameter. A uniform magnetic field of 2T in the horizontal direction exists such that initially the axis of the coil is in the direction of the field. The coil rotates through an angle of 90° under the influence of the magnetic field. (iii) What are the magnitudes of the torques on the coil in the initial and final position? (iv) What is the angular speed acquired by the coil when it has rotated by 90° ? The M.I. of the coil is 0.1 kg m².

Answer

(i) From Eq. (5.21)

$$B = \frac{\mu_0 Ni}{2r}$$

Here $N = 100$; $i = 3.2$ A, and $r = 0.1$ m. Hence,

$$\begin{aligned}B &= \frac{4\pi \times 10^{-7} \times 10^2 \times 3.2}{2 \times 10^{-1}} \\&= \frac{4 \times 10^{-5} \times 10}{2 \times 10^{-1}} \quad (\text{using } \pi \times 3.2 = 10) \\&= 2 \times 10^{-3} \text{ T}\end{aligned}$$

The direction is given by the right-hand thumb rule shown in Fig. 5.8.

(ii) The magnetic moment is given by Eq. (5.45),

$$\begin{aligned}m &= N i A \\&= N i \pi r^2 \\&= 100 \times 3.2 \times 3.14 \times 10^{-2} \\&= 10 \text{ A m}^2\end{aligned}$$

The direction is once again given by the right hand thumb rule.

(iii) $\tau = |\mathbf{m} \times \mathbf{B}|$ [from Eq. (5.44)]

$$= m B \sin \theta$$

Initially, $\theta = 0$. Thus, initial torque $\tau_i = 0$. Finally, $\theta = \pi/2$ (or 90°). Thus, final torque $\tau_f = m B = 10 \times 2 = 20 \text{ N m}$.

(iv) From Newton's second Law,

$$I \frac{d\omega}{dt} = m B \sin \theta$$

where I is the moment of inertia of the coil. From chain rule,

$$\frac{d\omega}{dt} = \frac{d\omega}{d\theta} \frac{d\theta}{dt} = \frac{d\omega}{d\theta} \omega$$

Using this,

$$I \omega d\omega = m B \sin \theta d\theta$$

Integrating from $\theta = 0$ to $\theta = \pi/2$,

$$I \int_0^{\omega_f} \omega d\omega = m B \int_0^{\pi/2} \sin \theta d\theta$$

$$\begin{aligned}I \frac{\omega_f^2}{2} &= -m B \cos \theta \Big|_0^{\pi/2} \\&= m B\end{aligned}$$

$$\begin{aligned}\omega_f &= \left(\frac{2mB}{I} \right)^{1/2} \\ &= \left(\frac{2 \times 20}{10^{-1}} \right)^{1/2} \\ &= 20 \text{ rad s}^{-1}.\end{aligned}$$

5.10 THE MOVING COIL GALVANOMETER (MCG)

Currents and voltages in circuits have been discussed extensively in Chapters 3 and 4. But how do we measure them? How do we claim that current in the circuit is 1.5 A or the voltage drop across a resistor is 1.2 V? Figure 5.26 exhibits a very useful instrument for this purpose : the moving coil galvanometer (MCG). It is a device whose principle can be understood on the basis of our discussion in Section 5.9.

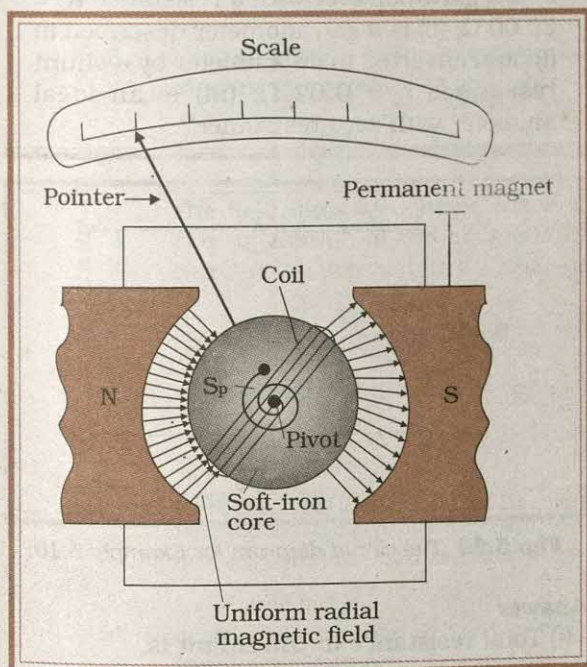


Fig. 5.26 The moving coil galvanometer. Its elements are described in the text. Depending on the requirement, this device can be used as a current detector or for measuring the value of the current (ammeter) or voltage (voltmeter).

The MCG consists of a multiturn coil free to rotate about a vertical axis (Fig. 5.26), in a uniform radial magnetic field. There is a cylindrical soft iron core to increase the

sensitivity of the MCG. When a current flows through the coil, a torque acts on it. This torque is given by Eqs. (5.44) and (5.45) to be

$$\tau = NiAB$$

where the symbols have their usual meaning. Since the field is radial by design, we have taken $\sin \theta \approx 1$ in the above expression for the torque. A spring S_p provides a counter torque that balances the magnetic torque resulting in a steady angular deflection ϕ . In equilibrium

$$k\phi = NiAB$$

where k is the torsional constant of the spring. It is the restoring torque per unit twist. The deflection ϕ is indicated on the scale by a pointer attached to the spring. We have

$$\phi = \left(\frac{NAB}{k} \right) i \quad (5.52)$$

The quantity in brackets is a constant for a given galvanometer.

The MCG can be used in a number of ways. It can be used as a detector to check if a current is flowing in the circuit. We have come across this usage in the Wheatstone's bridge arrangement. In this usage the neutral position of the pointer (when no current is flowing through the MCG) is in the middle of the scale and not at the left end as shown in Fig. 5.26. Depending on the direction of the current, the pointer deflection is either to the right or, the left.

The MCG can be used as an ammeter to measure the value of the current in a given circuit. For this the galvanometer must be connected in series with the circuit. The multiturn coil of the galvanometer has a resistance R_g and this will change the value of the current in the circuit. This is a classic case of how the introduction of a measuring device can affect the value of the measurement. To minimise this, one attaches a small resistance r_s , called the *shunt resistance*, in parallel with the galvanometer. The resistance of this arrangement is,

$$\frac{R_g r_s}{R_g + r_s} \approx r_s : \text{very small}$$

This arrangement is schematically shown in Fig. 5.27.

The scale of this ammeter is calibrated and then graduated to read off the current value with ease. We define the *sensitivity of the ammeter as the deflection per unit current*. From Eq. (5.52) this

current sensitivity is,

$$\frac{\phi}{i} = \frac{NAB}{k} \quad (5.53)$$

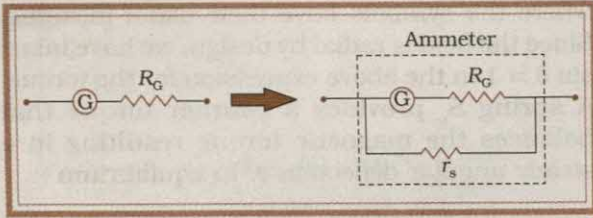


Fig. 5.27 Conversion of a galvanometer (G) to an ammeter (A) by the introduction of a shunt resistance r_s of very small value in parallel.

Note that such a definition of sensitivity is not arbitrary. For any arrangement, sensitivity is the ratio of the response to the stimulus. In our case, the response is the angular deflection and the stimulus is the current. A convenient way to increase the sensitivity is to increase the number of turns N .

The MCG can also be used as a voltmeter to measure the voltage across a given section of the circuit. For this it must be connected in parallel with that section of the circuit. Further, it must not draw current, otherwise the voltage measurement will not be accurate. To ensure accuracy, a large resistance R is attached in series with the MCG. This arrangement is schematically depicted in Fig. 5.28. Note that the resistance of the voltmeter is now,

$$R_G + R \approx R: \text{ large}$$

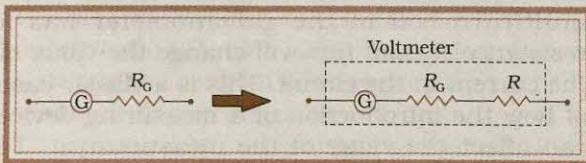


Fig. 5.28 Conversion of a galvanometer (G) to an voltmeter (V) by the introduction of a resistance R of large value in series.

The scale of the voltmeter is calibrated and then graduated to read off the voltage value with ease. We define the *voltage sensitivity* as the deflection per unit voltage. From Eq. (5.52),

$$\frac{\phi}{V} = \left(\frac{NAB}{k} \right) \frac{i}{V} = \left(\frac{NAB}{k} \right) \frac{1}{R} \quad (5.54)$$

An interesting point to note is that increasing the current sensitivity may not necessarily

increase the voltage sensitivity. Let us take Eq. (5.53) which provides a measure of current sensitivity. If $N \rightarrow 2N$, i.e., we double the number of turns then

$$\frac{\phi}{i} \rightarrow 2 \frac{\phi}{i}$$

Thus, the current sensitivity doubles. However, the resistance of the MCG is also likely to double, since it is proportional to the length of the wire. In Eq. (5.54), $N \rightarrow 2N$, and $R \rightarrow 2R$, thus the voltage sensitivity,

$$\frac{\phi}{V} \rightarrow \frac{\phi}{V}$$

remains unchanged.

Example 5.10 In the circuit shown below the current is to be measured. What is the value of the current if the ammeter shown (i) is a galvanometer with a resistance $R_G = 60.00 \, \Omega$; (ii) is a galvanometer described in (i) but converted to an ammeter by a shunt resistance $r_s = 0.02 \, \Omega$; (iii) is an ideal ammeter with zero resistance?

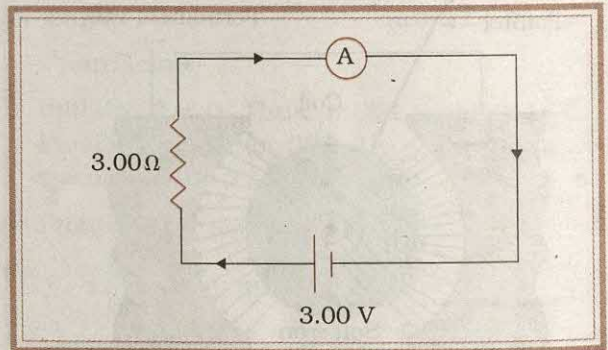


Fig. 5.29 The circuit diagram for Example 5.10.

Answer

(i) Total resistance in the circuit is,

$$R_G + 3 = 63 \, \Omega$$

$$\text{Hence, } i = \frac{3}{63} = 0.048 \text{ A}$$

(ii) Resistance of the galvanometer converted to an ammeter is,

$$\frac{R_G r_s}{R_G + r_s} = \frac{60 \times 0.02}{(60 + 0.02)} \approx 0.02 \, \Omega$$

(iii) Total resistance in the circuit is,

$$0.02 \, \Omega + 3 \, \Omega = 3.02 \, \Omega$$

Hence

$$i = \frac{3}{3.02} = 0.99 \text{ A}$$

(iv) For the ideal ammeter with zero resistance,

$$i = \frac{3}{3} = 1.00 \text{ A}$$

SUMMARY

1. The **Biot-Savart** law asserts that the magnetic field $d\mathbf{B}$ due to an element $d\mathbf{l}$ carrying a steady current i at a point P at a distance r from the current element is:

$$d\mathbf{B} = \frac{\mu_0}{4\pi} i \frac{d\mathbf{l} \times \mathbf{r}}{r^3}$$

Study Fig. 5.2 carefully. To obtain the total field at P, we must integrate this vector expression over the entire length of the conductor.

2. The constant μ_0 has the *exact* value

$$\mu_0 = 4\pi \times 10^{-7} \text{ T m A}^{-1}$$

It is the permeability of free space. It is related to ϵ_0 , the permittivity of free space and the speed of light in free space c , by

$$c = \frac{1}{\sqrt{\epsilon_0 \mu_0}}$$

3. The magnitude of the magnetic field at a distance R from a long, straight wire carrying a current i is given by:

$$B = \frac{\mu_0 i}{2\pi R}$$

The field lines are circles concentric with the wire.

4. The magnitude of the magnetic field due to a N -turn circular coil of radius R carrying a current i at an axial distance x from the centre is

$$B = \frac{\mu_0 N i R^2}{2(x^2 + R^2)^{3/2}}$$

At the center this reduces to

$$B = \frac{\mu_0 N i}{2R}$$

5. **Ampere's Circuital Law:** We have discussed a simplified form of this law. If \mathbf{B} is directed along the tangent to every point on the perimeter L of a closed curve and is constant in magnitude along perimeter then,

$$BL = \mu_0 i_e$$

where i_e is the net current enclosed by the closed circuit.

6. Employing Ampere's law one can show that the magnitude of the field B inside a long solenoid carrying a current i is

$$B = \mu_0 n i$$

where n is the number of turns per unit length. For a *toroid* one obtains,

$$B = \frac{\mu_0 N i}{2\pi r}$$

where N is the total number of turns and r the average radius.

7. The total force on a charge q moving with velocity \mathbf{v} in the presence of magnetic and electric fields \mathbf{B} and \mathbf{E} , respectively is called the *Lorentz force*. It is given by the expression:

$$\mathbf{F} = q(\mathbf{v} \times \mathbf{B} + \mathbf{E})$$

The magnetic force $q(\mathbf{v} \times \mathbf{B})$ is normal to \mathbf{v} and work done by it is zero.

8. In a uniform magnetic field \mathbf{B} , the charge q described in (7) executes circular orbit in the plane normal to \mathbf{B} . Its frequency of uniform circular motion is called the *cyclotron frequency* and is given by:

$$f_c = \frac{qB}{2\pi m}$$

This frequency is independent of the particle's speed and radius. This fact is exploited in a machine, the cyclotron, which is used to accelerate charged particles.

9. A straight conductor of length l and carrying a steady current i experiences a force \mathbf{F} in a uniform external magnetic field \mathbf{B} ,

$$\mathbf{F} = i\mathbf{l} \times \mathbf{B}$$

where $|\mathbf{l}| = l$ and the direction of \mathbf{l} is given by the direction of the current.

10. Parallel currents attract and anti-parallel currents repel.
11. A planar loop carrying a current i , having N closely, wound turns, and an area A possesses a magnetic moment \mathbf{m} where,

$$\mathbf{m} = N i \mathbf{A}$$

and the direction of \mathbf{m} is given by the right-hand thumb rule : curl the palm of your right hand along the loop with the fingers pointing in the direction of the current. The thumb sticking out gives the direction of \mathbf{m} (and \mathbf{A})

When this loop is placed in a uniform magnetic field \mathbf{B} , the force \mathbf{F} on it is

$$\mathbf{F} = 0$$

And the torque on it is,

$$\boldsymbol{\tau} = \mathbf{m} \times \mathbf{B}$$

In a moving coil galvanometer, this torque is balanced by a counter-torque due to a spring, yielding

$$k\phi = NiAB$$

where ϕ is the equilibrium deflection and k the torsion constant of the spring.

12. An electron moving around the central nucleus has a magnetic moment μ_l given by:

$$\mu_l = \frac{e}{2m} l$$

where l is the magnitude of the angular momentum of the circulating electron about the central nucleus. The smallest value of μ_l is called the Bohr magneton μ_B and it is

$$\mu_B = 9.27 \times 10^{-24} \text{ J/T}$$

Physical Quantity	Symbol	Nature	Dimensions	Units	Remarks
Permeability of free space	μ_0	Scalar	$[\text{MLT}^{-2}\text{A}^{-2}]$	T m A^{-1}	$4\pi \times 10^{-7} \text{ T m A}^{-1}$
Magnetic Field	\mathbf{B}	Vector	$[\text{M T}^{-2}\text{A}^{-1}]$	T (telsa)	
Magnetic Moment	\mathbf{m}	Vector	$[\text{L}^2\text{A}]$	A m^2 or J/T	
Torsion Constant	k	Scalar	$[\text{M L}^2\text{T}^{-2}]$	N m rad^{-1}	Appears in MCG

POINTS TO PONDER

1. Unlike the gravitational and the electrostatic fields, there is no scalar potential associated with the magnetic field.
2. $\mu_0 = 4\pi \times 10^{-7} \text{ TmA}^{-1}$ is an exact number and not an empirically obtained constant. It serves to define the ampere, the SI unit of current [via Eq. (5.40)] which in turn defines the coulomb.
3. Electrostatic field lines originate at a positive charge and terminate at a negative charge or fade at infinity. Magnetic field lines form closed loops. You may picture a field line to be like a snake which coils on itself with its tail in its mouth.
4. The discussion in this Chapter holds only for steady currents which do not fluctuate with time.
5. Recall the expression for the Lorentz force,

$$\mathbf{F} = q(\mathbf{v} \times \mathbf{B} + \mathbf{E})$$

This velocity dependent force has occupied the attention of some of the greatest scientific thinkers. If one switches to a frame with instantaneous velocity \mathbf{v} , the magnetic part of the force vanishes. The motion of the charged particle is then explained by arguing that there exists an appropriate electric field in the new frame. We shall not discuss the details of this mechanism – it is best left to advanced texts. However, we stress that the resolution of this paradox implies that electricity and magnetism are linked phenomena (*electromagnetism*) and that the Lorentz force expression *does not* imply a universal preferred frame of reference in nature.

6. Ampere's Circuital law is not independent of the Biot-Savart law. It can be derived from the Biot-Savart law. Its relationship to the Biot-Savart law is similar to the relationship between Gauss' law and Coulomb's law.
7. Why do we need a particle accelerating machine like the cyclotron to study the nucleus? Some insight can be obtained if we review the following facts. Raising the temperature of a gas by 50 K ensures observable changes in volume and pressure. This corresponds to a 5 meV (millielectron-volt) increase in the thermal energy of gas molecules. The cohesive energy of a solid is typically 1eV/atom. The bond strength of H_2 molecule is approximately 5 eV. The ionization energy of atoms is 10 to 50 eV. *The smaller the entity which is to be probed, the larger the energy needed.* For nuclear probes, we need energies of the order of 10 MeV. Current particle accelerators are aiming at TeV (10^{12} eV) energies in order to probe the structure of the elementary particles such as protons and electrons.
8. In section 5.8, we demonstrated that for parallel conductors and steady currents the Biot-Savart law and Lorentz force yield results in accordance with Newton's third Law. We stress this since Newton's third Law is not valid for electromagnetic phenomena in general. We shall, however, not address this issue in the present text-book.

EXERCISES

- 5.1 State the Biot-Savart law for the magnetic field due to a current-carrying element. Use this law to obtain a formula for magnetic field at the centre of a circular loop of radius a carrying a steady current I .
- 5.2 A circular coil of wire consisting of 100 turns, each of radius 8.0 cm carries a current of 0.40 A. What is the magnitude of the magnetic field \mathbf{B} at the centre of the coil?
- 5.3 Give the formula for the magnetic field produced by a straight infinitely long current-carrying wire. Describe the lines of field \mathbf{B} in this case.

- 5.4** A long straight wire carries a current of 35 A. What is the magnitude of the field \mathbf{B} at a point 20 cm from the wire?
- 5.5** A long straight wire in the horizontal plane carries a current of 50 A in north to south direction. Give the magnitude and direction of \mathbf{B} at a point 2.5 m east of the wire.
- 5.6** A horizontal overhead power line carries a current of 90 A in east to west direction. What is the magnitude and direction of the magnetic field due to the current 1.5 m below the line?
- 5.7** A straight wire carrying a current of 12 A is bent into a semi-circular arc of radius 2.0 cm as shown in Fig. 5.30(a). What is the direction and magnitude of \mathbf{B} at the centre of the arc? Would your answer be different if the wire were bent into a semi-circular arc of the same radius but in the opposite way as shown in Fig. 5.30(b)?

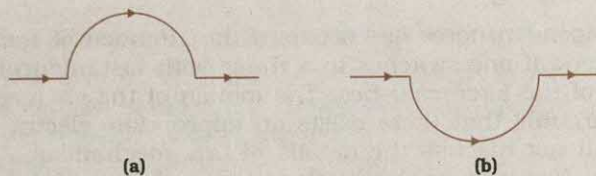


Fig. 5.30

- 5.8** What is the magnitude of magnetic force per unit length on a wire carrying a current of 8 A and making an angle of 30° with the direction of a uniform magnetic field of 0.15 T?
- 5.9** A 3.0 cm wire carrying a current of 10 A is placed inside a solenoid perpendicular to its axis. The magnetic field inside the solenoid is given to be 0.27 T. What is the magnetic force on the wire?
- 5.10** Two long and parallel straight wires A and B carrying currents of 8.0 A and 5.0 A in the same direction are separated by a distance of 4.0 cm. Estimate the force on a 10 cm section of wire A.
- 5.11** A closely wound solenoid 80 cm long has 5 layers of windings of 400 turns each. The diameter of the solenoid is 1.8 cm. If the current carried is 8.0 A, estimate the magnitude of \mathbf{B} inside the solenoid near its centre.
- 5.12** A square coil of side 10 cm consists of 20 turns and carries a current of 12 A. The coil is suspended vertically and the normal to the plane of the coil makes an angle of 30° with the direction of a uniform horizontal magnetic field of magnitude 0.80 T. What is the magnitude of torque experienced by the coil?
- 5.13** Two moving coil meters, M_1 and M_2 have the following particulars:
- $$R_1 = 10 \, \Omega, \quad N_1 = 30,$$
- $$A_1 = 3.6 \times 10^{-3} \, \text{m}^2, \quad B_1 = 0.25 \, \text{T}$$
- $$R_2 = 14 \, \Omega, \quad N_2 = 42,$$
- $$A_2 = 1.8 \times 10^{-3} \, \text{m}^2, \quad B_2 = 0.50 \, \text{T}$$
- (The spring constants are identical for the two meters). Determine the ratio of (a) current sensitivity and (b) voltage sensitivity of M_2 and M_1 .
- 5.14** In a chamber, a uniform magnetic field of 6.5 G ($1 \, \text{G} = 10^{-4} \, \text{T}$) is maintained. An electron is shot into the field with a speed of $4.8 \times 10^6 \, \text{m s}^{-1}$ normal to the field. Explain why the path of the electron is a circle. Determine the radius of the circular orbit. ($e = 1.6 \times 10^{-19} \, \text{C}$, $m_e = 9.1 \times 10^{-31} \, \text{kg}$).

- 5.15** In Exercise 5.14 obtain the frequency of revolution of the electron in its circular orbit. Does the answer depend on the speed of the electron? Explain.
- 5.16** (a) A circular coil of 30 turns and radius 8.0 cm carrying a current of 6.0 A is suspended vertically in a uniform horizontal magnetic field of magnitude 1.0 T. The field lines make an angle of 60° with the normal of the coil. Calculate the magnitude of the counter torque that must be applied to prevent the coil from turning.
- (b) Would your answer change, if the circular coil in (a) were replaced by a planar coil of some irregular shape that encloses the same area? (All other particulars are also unaltered.)

ADDITIONAL EXERCISE

- 5.17** Two concentric circular coils X and Y of radii 16 cm and 10 cm, respectively, lie in the same vertical plane containing the north to south direction. Coil X has 20 turns and carries a current of 16 A; coil Y has 25 turns and carries a current of 18 A. The sense of the current in X is anticlockwise, and clockwise in Y, for an observer looking at the coils facing west. Give the magnitude and direction of the net magnetic field due to the coils at their centre.
- 5.18** A magnetic field of 100 G ($1 \text{ G} = 10^{-4} \text{ T}$) is required which is uniform in a region of linear dimension about 10 cm and area of cross-section about 10^{-3} m^2 . The maximum current-carrying capacity of a given coil of wire is 15 A and the number of turns per unit length that can be wound round a core is at most 1000 turns m^{-1} . Suggest some appropriate design particulars of a solenoid for the required purpose. Assume the core is not ferromagnetic.
- 5.19** For a circular coil of radius R and N turns carrying current I , the magnitude of the magnetic field at a point on its axis at a distance x from its centre is given by,

$$B = \frac{\mu_0 I R^2 N}{2(x^2 + R^2)^{3/2}}$$

- (a) Show that this reduces to the familiar result for field at the centre of the coil.
- (b) Consider two parallel co-axial circular coils of equal radius R , and number of turns N , carrying equal currents in the same direction, and separated by a distance R . Show that the field on the axis around the mid-point between the coils is uniform over a distance that is small as compared to R , and is given by,

$$B = 0.72 \frac{\mu_0 N I}{R}, \text{ approximately.}$$

[Such an arrangement to produce a nearly uniform magnetic field over a small region is known as *Helmholtz coils*.]

- 5.20** A toroid has a core (non-ferromagnetic) of inner radius 25 cm and outer radius 26 cm, around which 3500 turns of a wire are wound. If the current in the wire is 11 A, what is the magnetic field (i) outside the toroid, (ii) inside the core of the toroid, and (iii) in the empty space surrounded by the toroid.
- 5.21** Answer the following questions:
- (a) A magnetic field that varies in magnitude from point to point but has a

constant direction (east to west) is set up in a chamber. A charged particle enters the chamber and travels undeflected along a straight path with constant speed. What can you say about the initial velocity of the particle?

- (b) A charged particle enters an environment of a strong and non-uniform magnetic field varying from point to point both in magnitude and direction, and comes out of it following a complicated trajectory. Would its final speed equal the initial speed if it suffered no collisions with the environment?
- (c) An electron travelling west to east enters a chamber having a uniform electrostatic field in north to south direction. Specify the direction in which a uniform magnetic field should be set up to prevent the electron from deflecting from its straight line path.

5.22 Answer the following questions:

- (a) A cloud chamber photograph shows a pair of circular tracks emerging from a common point. The tracks have similar density at droplets but curve in opposite directions in a plane normal to the magnetic field maintained in the chamber. If one of the ionising particles is established to be an electron, guess the high energy event that took place at the common point of the tracks.
- (b) A similar event as in (a) photographed in a liquid-hydrogen bubble chamber shows spiral tracks instead of circular tracks. Explain why?

5.23 An electron emitted by a heated cathode and accelerated through a potential difference of 2.0 kV, enters a region with uniform magnetic field of 0.15 T. Determine the trajectory of the electron if the field (a) is transverse to its initial velocity, (b) makes an angle of 30° with the initial velocity.

5.24 A cyclotron's oscillator frequency is 10 MHz. What should be the operating magnetic field for accelerating protons? If the radius of its 'dees' is 60 cm, what is the kinetic energy of the proton beam produced by the accelerator? ($e = 1.60 \times 10^{-19}$ C, $m_p = 1.67 \times 10^{-27}$ kg). Express your answer in units of MeV (1 MeV = 1.602×10^{-13} J).

5.25 A magnetic field set up using Helmholtz coils (described in Exercise 5.19) is uniform in a small region and has a magnitude of 0.75 T. In the same region, a uniform electrostatic field is maintained in a direction normal to the common axis of the coils. A narrow beam of (single species) charged particles all accelerated through 15 kV enters this region in a direction perpendicular to both the axis of the coils and the electrostatic field. If the beam remains undeflected when the electrostatic field is 9.0×10^5 V m $^{-1}$, make a simple guess as to what the beam contains. Why is the answer not unique?

5.26 A straight horizontal conducting rod of length 0.45 m and mass 60 g is suspended by two vertical wires at its ends. A current of 5.0 A is set up in the rod through the wires.

- (a) What magnetic field should be set up normal to the conductor in order that the tension in the wires is zero?
- (b) What will be the total tension in the wires if the direction of current is reversed keeping the magnetic field same as before? (Ignore the mass of the wires.) $g = 9.8$ m s $^{-2}$.

5.27 The wires which connect the battery of an automobile to its starting motor carry a current of 300 A (for a short time). What is the force per unit length between the wires if they are 70 cm long and 1.5 cm apart? Is the force attractive or repulsive?

- 5.28** A uniform magnetic field of 1.5 T exists in a cylindrical region of radius 10.0 cm, its direction parallel to the axis along east to west. A wire carrying current of 7.0 A in the north to south direction passes through this region. What is the magnitude and direction of the force on the wire if,
- the wire intersects the axis,
 - the wire is turned from N-S to northeast-northwest direction,
 - the wire in the N-S direction is lowered from the axis by a distance of 6.0 cm?

- 5.29** A uniform magnetic field of 3000 G is established along the positive z-direction. A rectangular loop of sides 10 cm and 5 cm carries a current of 12 A. What is the torque on the loop in the different cases shown in the Fig. 5.31? What is the force on each case? Which case corresponds to stable equilibrium?

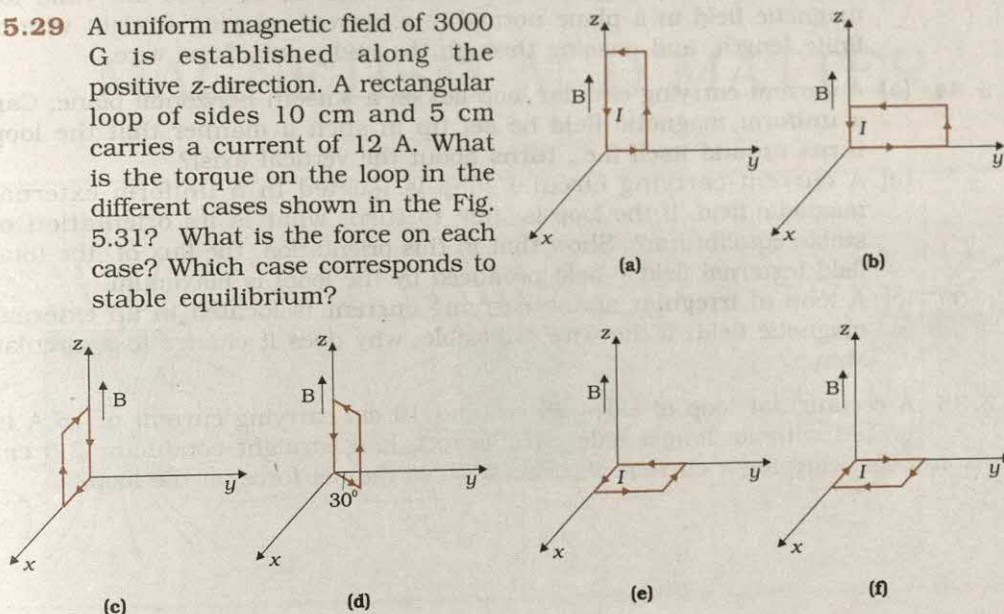


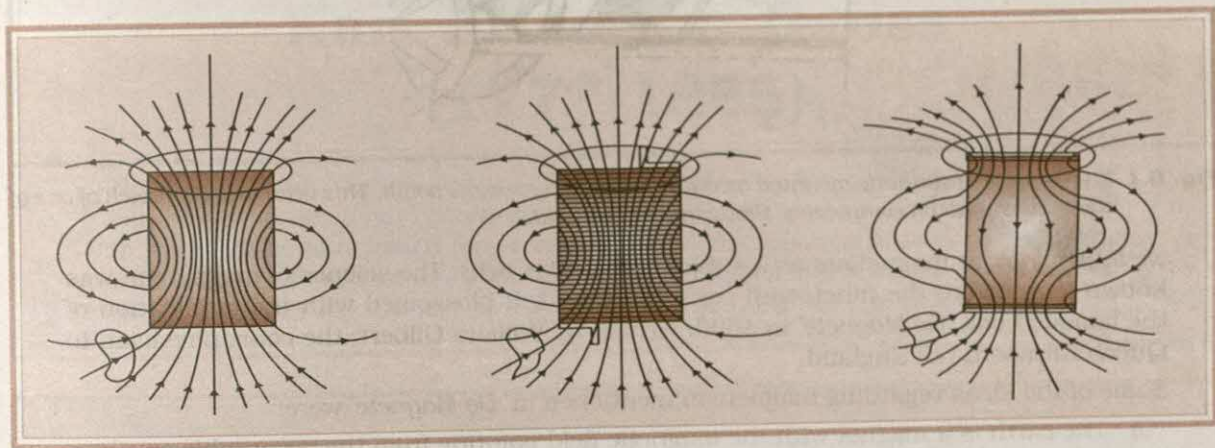
Fig. 5.31

- 5.30** A circular coil of 20 turns and radius 10 cm is placed in a uniform magnetic field of 0.10 T normal to the plane of the coil. If the current in the coil is 5.0 A, what is the
- total torque on the coil,
 - total force on the coil,
 - average force on each electron in the coil due to the magnetic field?
- (The coil is made of copper wire of cross-sectional area 10^{-5} m^2 , and the free electron density in copper is given to be about 10^{29} m^{-3} .)
- 5.31** A solenoid 60 cm long and of radius 4.0 cm has 3 layers of windings of 300 turns each. A 2.0 cm long wire of mass 2.5 g lies inside the solenoid (near its centre) normal to its axis; both the wire and the axis of the solenoid are in the horizontal plane. The wire is connected through two leads parallel to the axis of the solenoid to an external battery which supplies a current of 6.0 A in the wire. What value of current (with appropriate sense of circulation) in the windings of the solenoid can support the weight of the wire? $g = 9.8 \text{ m s}^{-2}$.
- 5.32** Describe qualitatively the path of a charged particle moving in
- a uniform electrostatic field, with initial velocity (i) parallel to the field, (ii) perpendicular to the field, (iii) at an arbitrary angle with the field direction;
 - a uniform magnetic field, with initial velocity (i) parallel to the field, (ii) perpendicular to the field, (iii) at an arbitrary angle with the field direction;

- (c) a region with uniform electrostatic and magnetic fields parallel to each other, with initial velocity (i) parallel, (ii) perpendicular, (iii) at an arbitrary angle with the common direction of the fields.
- 5.33** (a) State Ampere's law connecting the line integral of \mathbf{B} over a closed path to the net current crossing the area bounded by the path.
 (b) Use Ampere's law to derive the formula for magnetic field due to an infinitely long straight current-carrying wire.
 (c) Explain carefully why the same derivation as in (b) is not valid for magnetic field in a plane normal to a current-carrying straight wire of finite length, and passing through the mid-point of the wire.
- 5.34** (a) A current-carrying circular loop lies on a smooth horizontal plane. Can a uniform magnetic field be set up in such a manner that the loop turns around itself (i.e., turns about the vertical axis)?
 (b) A current-carrying circular loop is located in a uniform external magnetic field. If the loop is free to turn, what is its orientation of stable equilibrium? Show that in this orientation, the flux of the total field (external field + field produced by the loop) is maximum.
 (c) A loop of irregular shape carrying current is located in an external magnetic field. If the wire is flexible, why does it change to a circular shape?
- 5.35** A rectangular loop of sides 25 cm and 10 cm carrying current of 15 A is placed with its longer side parallel to a long straight conductor 2.0 cm apart carrying a current of 25 A. What is the net force on the loop?

CHAPTER SIX

MAGNETISM AND MATTER



6.1 INTRODUCTION

Magnetic phenomena are universal in nature. Vast, distant galaxies and atoms, man and beast, are permeated through and through with a host of magnetic fields from a variety of sources. The earth's magnetism predates human evolution. The word magnet is derived from the name of an island in Greece called 'magnesia' where magnetic ore deposits as early as 800 BC were found. Shepherds on this island complained that their wooden shoes (which had nails) at times stayed stuck to the ground. Their iron-tipped rods were similarly affected. This attractive property of magnets made it difficult for them to move around.

The directional property of magnets was known since ancient times. A thin long piece of a magnet, when suspended freely, pointed in the north-south direction. A similar effect was observed when it was placed on a piece of cork which was then allowed to float in still water. The name lodestone which is given to magnetite means 'leading stone'. The technological exploitation of this property is generally credited to the Chinese. Chinese texts dating 400 BC mention the use of magnetic needles for navigation on ships. Caravans crossing the Gobi desert also employed magnetic needles.

A Chinese legend narrates the tale of the victory of the emperor Huang-ti about four thousand years ago, which he owed to his craftsmen (whom nowadays you would call engineers). These 'engineers' built a chariot on which they placed a magnetic figure with arms outstretched. Figure 6.1 is an artist's description of this chariot. The figure swivelled around so that the finger always pointed south. With this chariot, Huang-ti's troops were able to attack the enemy from the rear in thick fog, and to defeat them.

In the previous chapter we had learnt that moving charges or currents produced magnetic fields. This discovery, which was made in the early part of the nineteenth century is credited to Oersted, Ampere, and Biot and Savart, among others. In the present chapter,

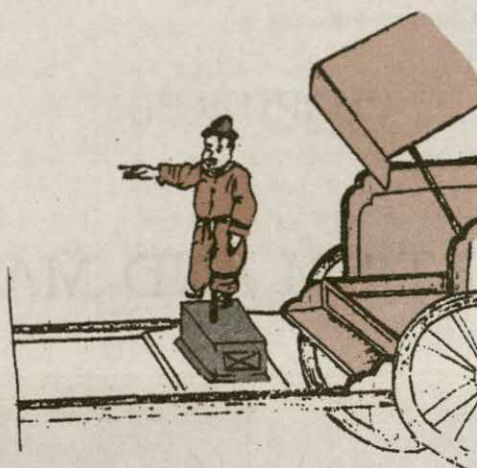


Fig. 6.1 The arm of the statuette mounted on the chariot always points south. This is an artist's sketch of one of the earliest known compasses, thousands of year old.

we take a look at magnetism as a subject in its own right. The science of magnetism was known long before the nineteenth century. Indeed, it blossomed with the publication of the famous book '*De Magnete*' in 1600, written by William Gilbert, the court physician to Queen Elizabeth I of England.

Some of the ideas regarding magnetism mentioned in '*De Magnete*' were:

- (i) The earth is a magnet with the magnetic field pointing from the geographic south to the north. It probably consists of a 'giant bar magnet' placed approximately along its axis of rotation*. We note that the word 'field' in the term magnetic field is of later origin.
- (ii) When a bar magnet is freely suspended, or floated in still water, it points in the north-south direction. The tip which points to the geographic north is called the north pole, and the tip which points to the geographic south is called the south pole.
- (iii) There is a repulsive force when the two north poles (or the two south poles) of two magnets are brought close together. Conversely, there is an attractive force between the north pole of one bar magnet and the south pole of the other.
- (iv) We cannot isolate the north, or south pole of a magnet. If a bar magnet is sliced into half, we get two smaller bar magnets with somewhat weaker directional properties. Unlike electric charges, *magnetic monopoles do not exist*.
- (v) It is possible to make magnets out of iron and its alloys. This technological accomplishment was known prior to 1800 and will be described later.

Gilbert's book laid the foundations for a scientific approach to magnetic phenomena and to terrestrial magnetism. In this chapter, we begin with a description of the bar magnet and its behaviour in an external magnetic field. We describe Gauss's law for magnetism. We then follow it up with an account of the earth's magnetic field. We next describe how materials can be classified on the basis of their magnetic properties. We describe para-, dia-, and ferromagnetism. We conclude with a section on electromagnets and permanent magnets.

* The earth's core is at high temperature. It is known that materials lose their magnetic property at high temperatures. So this hypothesis is wrong.



Karl Friedrich Gauss (1777-1855)

He was a child prodigy and was gifted in mathematics, physics, engineering, astronomy and even land surveying. The properties of numbers fascinated him, and in his work he anticipated major mathematical development of later times. Along with Wilhelm Welser, he built the first electric telegraph in 1833. His mathematical theory of curved surface laid the foundation for the later work of Riemann.



Hendrik Antoon Lorentz (1853-1928)

Dutch theoretical physicist, professor at Leiden. He investigated the relationship between electricity, magnetism, and mechanics. In order to explain the observed effect of magnetic fields on emitters of light (Zeeman effect), he postulated the existence of electric charges in the atom, for which he was awarded the Nobel Prize in 1902. He derived a set of transformation equations (known after him, as Lorentz transformation equations) by some tangled mathematical arguments, but he was not aware that these equations hinge on a new concept of space and time.

6.2 THE BAR MAGNET

One of the earliest childhood memories of the famous physicist Albert Einstein was that of a magnet gifted to him by a relative. Einstein was fascinated, and played endlessly with it. He wondered how the magnet could affect objects such as nails or pins placed away from it and not in any way 'connected' to it by a spring or string.



Fig. 6.2 The arrangement of iron filings surrounding a bar magnet. The pattern mimics magnetic field lines. They suggest that the bar magnet is a magnetic dipole.

We begin our study by examining iron filings sprinkled on a sheet of glass placed over a short bar magnet. The arrangement of iron filings is shown in Fig. 6.2.

The pattern of iron filings suggests that the magnet has two poles similar to the positive and negative charge of an electric dipole. As mentioned in the introductory section, one pole is designated the north pole and the other, the south pole. When suspended freely, these poles point approximately towards the geographic north and south poles, respectively. A similar pattern of iron filings is observed around a current carrying solenoid.

6.2.1 The Magnetic Field Lines

The pattern of iron filings permits us to plot the magnetic field lines*. This is shown both for the bar magnet and the current-carrying solenoid in Fig. 6.3. For comparison the electric lines of force of an electric dipole are also displayed. The electric field lines are discussed in Chapter 1. The magnetic field lines are a visual and intuitive realization of the 'unseen' magnetic field. Their properties are:

- (i) The magnetic field lines of a magnet (or a solenoid) form continuous closed loops. This is unlike the electric dipole where

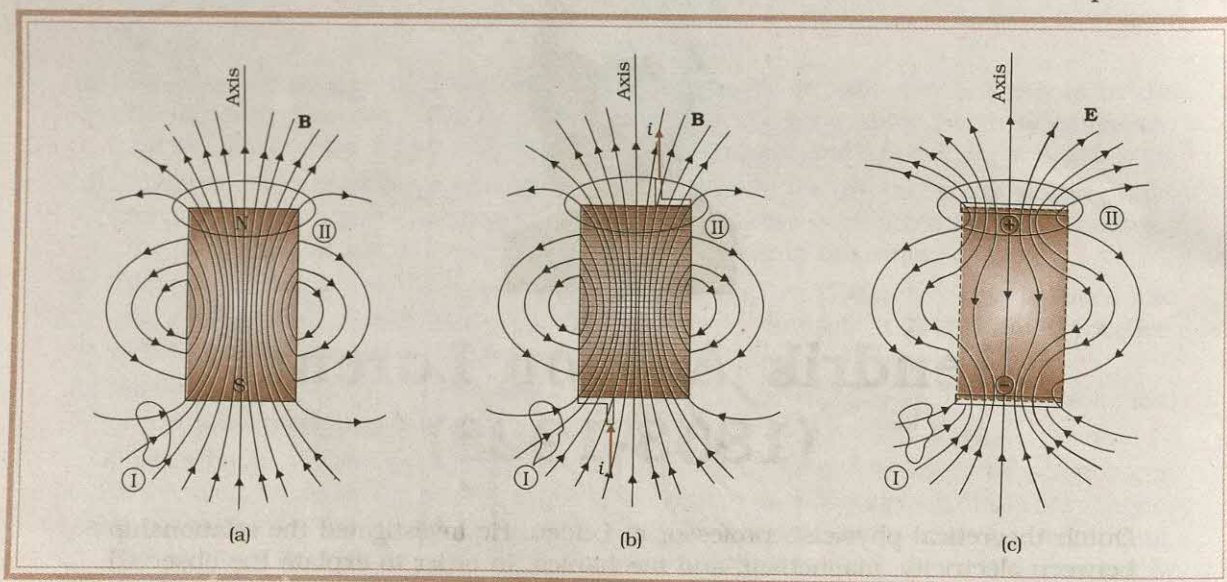


Fig. 6.3 The field lines of (a) a bar magnet, (b) a current-carrying finite solenoid, and (c) an electric dipole. At large distances, the three lines are very similar. The curves labelled I and II are closed Gaussian surfaces.

* In some text books the magnetic field lines are called 'magnetic lines of force'. This nomenclature is avoided since it can be confusing. Unlike electrostatics the field lines in magnetism do not indicate the direction of the force on a (moving) charge.

these lines begin on a positive charge and end on the negative charge [Fig. 6.3(c)].

- (ii) The tangent to the field line at a given point represents the direction of the net magnetic field \mathbf{B} at that point.
- (iii) The larger the number of field lines crossing per unit normal area, the larger is the magnitude of the magnetic field \mathbf{B} . In Fig. 6.3(a), \mathbf{B} is larger around region II than in region I.
- (iv) The magnetic field lines do not intersect. This is so since the direction of the magnetic field would not be unique at the point of intersection.

One can plot the magnetic field lines in a variety of ways. One way is to place a small magnetic compass needle at various positions and note its orientation. This gives us an idea of the magnetic field direction at various points in space.

6.2.2 The Bar Magnet as an Equivalent Solenoid

In the previous chapter, we have explained how a current loop acts as a magnetic dipole (Section 5.9). We mentioned Ampere's hypothesis that all magnetic phenomena can be explained in terms of circulating currents. Recall that the magnetic dipole moment \mathbf{m} associated with a current loop was defined to be $\mathbf{m} = Ni\mathbf{A}$ where N is the number of turns in the loop, i the current and \mathbf{A} the area vector [Eq. (5.45)].

The resemblance of magnetic field lines for a bar magnet and a solenoid suggest that a bar magnet may be thought of as a large number of circulating currents in analogy with a solenoid. Cutting a bar magnet in half is like cutting a solenoid. We get two smaller solenoids with weaker magnetic properties. The field lines remain continuous, emerging from one face of the solenoid and entering from the other. One can test this analogy by moving a small compass needle in the neighbourhood of a bar magnet and a current carrying finite solenoid and noting that the deflections of the needle are similar in both cases.

To make this analogy more firm we calculate the axial field of a finite solenoid depicted in Fig. 6.4(a). We shall demonstrate that at large distances this axial field resembles that of a bar magnet.

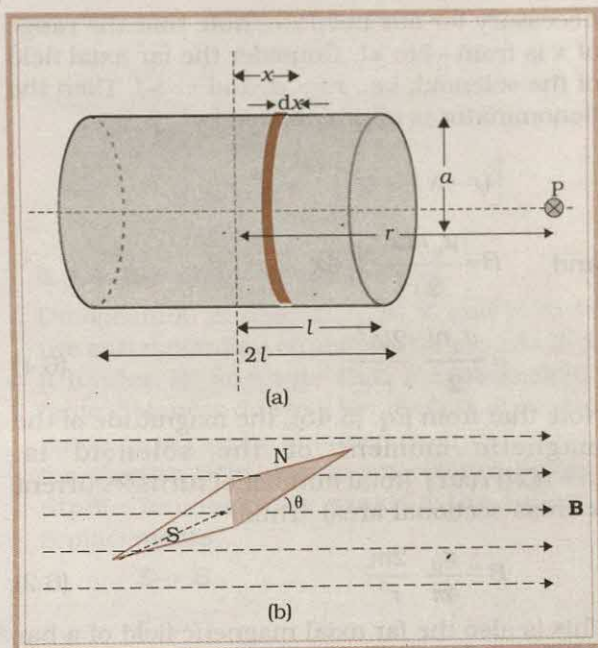


Fig. 6.4 Calculation of (a) The axial field of a finite solenoid in order to demonstrate its similarity to that of a bar magnet. (b) The time period of oscillation of a magnetic needle in a uniform magnetic field \mathbf{B} . The arrangement may be used to determine either \mathbf{B} or the magnetic moment \mathbf{m} of the needle.

Let the solenoid of Fig. 6.4(a) consist of n turns per unit length and of radius a . Let its length be $2l$. We can evaluate the axial field at a distance r from the centre of the solenoid (point P). To do this, consider a circular element dx of the solenoid at a distance x from its centre. It consists of ndx turns. Let i be the current in the solenoid. In Section 5.3.2 of the previous chapter we have calculated the magnetic field on the axis of a circular current loop. From Eq. (5.20), the magnitude of the field at point P due to the circular element is,

$$dB = \frac{\mu_0 n dx i a^2}{2[(r-x)^2 + a^2]^{3/2}}$$

The magnitude of the total field is obtained by summing over all the elements – in other words by integration from $x = -l$ to $x = +l$

$$B = \frac{\mu_0 n i a^2}{2} \int_{-l}^{+l} \frac{dx}{[(r-x)^2 + a^2]^{3/2}}$$

This integration can be done by trigonometric substitutions. This exercise, however, is not

necessary for our purpose. Note that the range of x is from $-l$ to $+l$. Consider the far axial field of the solenoid, i.e., $r \gg a$, and $r \gg l$. Then the denominator is approximated by

$$[(r-x)^2 + a^2]^{3/2} \approx r^3$$

$$\text{and } B = \frac{\mu_0 n i a^2}{2 r^3} \int_{-l}^l dx$$

$$\approx \frac{\mu_0 n i}{2} \frac{2la^2}{r^3} \quad (6.1)$$

Note that from Eq. (5.45), the magnitude of the magnetic moment of the solenoid is, $m = n(2l) i (\pi a^2)$ (total number of turns \times current \times cross-sectional area). Thus,

$$B \approx \frac{\mu_0}{4\pi} \cdot \frac{2m}{r^3} \quad (6.2)$$

This is also the far axial magnetic field of a bar magnet which one may obtain experimentally. Thus, a bar magnet and a solenoid produce similar magnetic field. The magnetic moment of a bar magnet is thus equal to the magnetic moment of an equivalent solenoid that produces the same magnetic field.

Some textbooks assign a 'magnetic charge' $^* +q_m$ to the north pole and $-q_m$ to the south pole of a bar magnet of length $2l$, and magnetic moment $q_m(2l)$. The field strength due to q_m at a distance r from it is given by $\mu_0 q_m / 4\pi r^2$. The magnetic field due to the bar magnet is then obtained, both for the axial and the equatorial case, in a manner analogous to that of an electric dipole (Chapter 1). The method is simple and appealing. However, magnetic monopoles do not exist, and we have avoided this approach for that reason.

6.2.3 The Dipole in a Uniform Magnetic Field

The pattern of iron filings and the magnetic field lines give us an approximate idea of the magnetic moment \mathbf{m} and its field \mathbf{B} . We may at times be required to determine the magnitude of \mathbf{B} accurately. This is done by placing a thin compass needle of known magnetic moment \mathbf{m} and moment of inertia I and allowing it to oscillate in the magnetic field. This arrangement is shown in Fig. 6.4(b).

The torque on the needle is [see Eq. (5.44)],

$$\tau = \mathbf{m} \times \mathbf{B} \quad (6.3)$$

$$\tau(\theta) = mB \sin \theta$$

From Newton's second law

$$I \frac{d^2 \theta}{dt^2} = -mB \sin \theta$$

$$\approx -mB \theta$$

where we approximate $\sin \theta \approx \theta$ for small values of θ and in radians.

Thus,

$$\frac{d^2 \theta}{dt^2} = -\frac{mB}{I} \theta$$

This represents simple harmonic motion. The square of the angular frequency $\omega^2 = mB/I$. The time period is,

$$T = 2\pi \sqrt{\frac{I}{mB}}$$

Example 6.1 In Fig. 6.4, the magnetic needle has magnetic moment $6.7 \times 10^{-2} \text{ Am}^2$ and moment of inertia $I = 7.5 \times 10^{-6} \text{ kg m}^2$. It performs 10 complete oscillations in 6.70 s. What is the magnitude of the magnetic field?

Answer The time period of oscillation is,

$$T = \frac{6.70}{10} = 0.67 \text{ s}$$

From the equation for time period derived before,

$$B = \frac{4\pi^2 I}{mT^2}$$

$$= \frac{4 \times (3.14)^2 \times 7.5 \times 10^{-6}}{6.7 \times 10^{-2} \times (0.67)^2}$$

$$= 0.01 \text{ T}$$

$$= 100 \text{ G}$$

[(Note 1 T (tesla) = 10^4 G (gauss)).]

The magnetic potential energy U_m is given by

$$U_m = \int \tau(\theta) d\theta$$

$$= \int mB \sin \theta$$

$$= -mB \cos \theta$$

$$= -\mathbf{m} \cdot \mathbf{B} \quad (6.4)$$

* Also called pole strength.

We have emphasized earlier that the zero of potential energy can be fixed at one's convenience. In Eq. (6.4), an additional constant can be introduced. Taking that constant to be zero means we have fixed the zero of potential energy at $\theta = 90^\circ$, when the needle is perpendicular to the field. Equation (6.4) shows that potential energy is minimum ($= -mB$) at $\theta = 0^\circ$ (most stable position) and maximum ($= +mB$) at $\theta = 180^\circ$ (most unstable position).

Example 6.2 A short bar magnet placed with its axis at 30° experiences a torque of 0.016 N m in an external field of 800 G . (a) What is the magnetic moment of the magnet? (b) What is the work done by an external force in moving it from its most stable to most unstable position? (c) What is the work done by the force due to the external magnetic field in the process mentioned in part (b)? (d) The bar magnet is replaced by a solenoid of cross-sectional area $2 \times 10^{-4} \text{ m}^2$ and 1000 turns, but the same magnetic moment. Determine the current flowing through the solenoid.

Answer

(a) From Eq. (6.3),

$$\tau = mB \sin \theta$$

$\theta = 30^\circ$, hence $\sin \theta = 1/2$. Thus

$$0.016 = m \times (800 \times 10^{-4}) \times \frac{1}{2}$$

$$m = \frac{160 \times 2}{800} = 0.40 \text{ Am}^2$$

(b) From Eq. (6.4), the most stable position is $\theta = 0^\circ$ and the most unstable position is $\theta = 180^\circ$. Work done by the external force is

$$W = U_m(\theta = 180^\circ) - U_m(\theta = 0^\circ)$$

$$= 2 m B$$

$$= 2 \times 0.40 \times 800 \times 10^{-4}$$

$$= 0.064 \text{ J}$$

(c) Here the displacement and the torque due to the magnetic field are in opposition. The work done by the force due to the external magnetic field is

$$W_B = -0.064 \text{ J}$$

(d) From Eq. (5.45)

$$m_s = NiA$$

From part (a), $m_s = 0.40 \text{ A m}^2$

$$0.40 = 1000 \times i \times 2 \times 10^{-4}$$

$$i = \frac{0.40 \times 10^4}{1000 \times 2} = 2 \text{ A}$$

6.2.4 The Electrostatic Analog

Composition of Eqs. (6.2), (6.3) and (6.4) with the corresponding equations for electric dipole (Chapter 1), suggests that magnetic field at large distances due to a bar magnet of magnetic moment \mathbf{m} can be obtained from the equation for electric field due to an electric dipole of dipole moments by making the following replacements:

$$\mathbf{E} \rightarrow \mathbf{B}$$

$$\mathbf{p} \rightarrow \mathbf{m}$$

$$\frac{1}{4\pi\epsilon_0} \rightarrow \frac{\mu_0}{4\pi}$$

In particular, we can write down the equatorial field (\mathbf{B}_E) of a bar magnet at a distance r , for $r \gg l$, where l is the size of the magnet:

$$\mathbf{B}_E = -\frac{\mu_0 \mathbf{m}}{4\pi r^3} \quad (6.5)$$

Likewise, the axial field (\mathbf{B}_A) of a bar magnet for $r \gg l$ is:

$$\mathbf{B}_A = \frac{2\mu_0 \mathbf{m}}{4\pi r^3} \quad (6.6)$$

Eq. (6.6) is just Eq. (6.2) in vector form. Table 6.1 summarises the analogy between electric and magnetic dipoles.

Table 6.1 The Dipole Analogy

	Electrostatics	Magnetism
	$1/\epsilon_0$	μ_0
Dipole moment	\mathbf{p}	\mathbf{m}
Equatorial Field	$-\mathbf{p}/4\pi\epsilon_0 r^3$	$-\mu_0 \mathbf{m}/4\pi r^3$
Axial Field	$2\mathbf{p}/4\pi\epsilon_0 r^3$	$\mu_0 2\mathbf{m}/4\pi r^3$
External Field: Torque	$\mathbf{p} \times \mathbf{E}$	$\mathbf{m} \times \mathbf{B}$
External Field: Energy	$-\mathbf{p} \cdot \mathbf{E}$	$-\mathbf{m} \cdot \mathbf{B}$

Example 6.3 What is the magnitude of the equatorial and axial fields due to a bar magnet of length 5 cm at a distance of 50 cm from its mid-point? The magnetic moment of the bar magnet is 0.40 A m^2 , the same as in Example 6.2.

Answer From Eq. (6.5)

$$\begin{aligned} B_E &= \frac{\mu_0 m}{4\pi r^3} \\ &= \frac{10^{-7} \times 0.4}{(0.5)^3} = \frac{10^{-7} \times 0.4}{0.125} \\ &= 3.2 \times 10^{-7} \text{ T, or } 3.2 \times 10^{-3} \text{ G} \\ B_A &= \frac{\mu_0 2m}{4\pi r^3} \\ &= 6.4 \times 10^{-7} \text{ T or } 6.4 \times 10^{-3} \text{ G} \end{aligned}$$

These fields are quite small, say in comparison to the earth's field of magnitude $\approx 0.4 \text{ G}$. From Example 6.2(d) we know that an equivalent solenoid of 1000 turns and 2 A current will reproduce this field due to a 5 cm bar magnet. Magnetic effects are indeed rather small. ◀

6.3 MAGNETISM AND GAUSS'S LAW

In Chapter 1, we studied Gauss's law for electrostatics. In Fig. 6.3(c), we see that for a closed surface represented by I, the number of lines leaving the surface is equal to the number of lines entering it. This is consistent with the fact that no net charge is enclosed by the surface. However, in the same figure, for the closer surface II there is a net outward flux, since it does include a net (positive) charge.

The situation is radically different for magnetic fields which are continuous and form closed loops. Examine the Gaussian surfaces represented by I or II in Fig. 6.3(a) or Fig. 6.3(b). Both cases visually demonstrate that the number of lines of force leaving the surface is balanced by the number entering it. The net magnetic flux is zero for both the surfaces. This is true for any closed surface.

Consider a small vector area element $\Delta \mathbf{S}$ of a closed surface S . The magnetic flux through $\Delta \mathbf{S}$ is defined as $\Delta \phi_B = \mathbf{B} \cdot \Delta \mathbf{S}$, where \mathbf{B} is the field at $\Delta \mathbf{S}$. We divide S into many small area elements and calculate the individual fluxes through each. Then, the net flux ϕ_B is,

$$\phi_B = \sum_{\text{all}} \Delta \phi_B = \sum_{\text{all}} \mathbf{B} \cdot \Delta \mathbf{S}$$

where 'all' stands for 'all area elements $\Delta \mathbf{S}$ '. Thus Gauss's law for magnetism is:

'The net magnetic flux through any closed surface is zero'.

Gauss's law is a reflection of the fact that isolated magnetic poles (also called *monopoles*) do not exist. There are no sources or sinks of \mathbf{B} ; the simplest magnetic element is a dipole or current loop. All magnetic phenomena can be explained in terms of an arrangement of dipoles and/or current loops.

6.4 THE EARTH'S MAGNETISM

We have referred to the magnetic field of the earth. Its value on the earth's surface is a few tenths of a gauss ($1 \text{ G} = 10^{-4} \text{ T}$). This magnetic field was thought of as arising from a gigantic bar magnet placed approximately along the axis of rotation of the earth and deep in its interior. This is shown in Fig. 6.5. We hasten to clarify that this is just a conceptual device adopted to explain the earth's magnetic field and its peculiarities. In reality there is no such bar magnet in the earth's interior.

Let us denote the geographic north and south poles by N_g and S_g , respectively. The magnetic axis of the earth makes an angle of approximately 20° with the geographic axis. Since the north pole of a compass needle points approximately to the geographic north N_g , we designate the earth's magnetic pole close to N_g as S_m , the south magnetic pole. Recall that opposite poles attract. Similarly, N_m , the north magnetic pole of the earth is close to S_g . These aspects are displayed in Fig. 6.5. S_m and N_m are points on the earth's surface, not in the interior. Hence we can specify them exactly in terms of latitude and longitude. S_m is located at a point in Northern Canada with latitude at 70.5°N and longitude at 96°W . N_m is located at a point diametrically opposite, 70.5°S 84°E . The magnetic poles are approximately 2000 km away from the geographic poles. The magnetic equator intersects the geographic equator at longitudes 6°W and 174°E , respectively.

Example 6.4 The earth's magnetic field at the equator is approximately 0.4 G. Estimate the earth's dipole moment.

Answer From Eq. (6.5), the equatorial magnetic field is,

$$B_E = \frac{\mu_0 m}{4\pi r^3}$$

We are given that $B_E \sim 0.4 \text{ G} (= 4 \times 10^{-5} \text{ T})$. For r , we take the radius of the earth $r = 6.4 \times 10^6 \text{ m}$. Hence,

$$\begin{aligned} m &= \frac{4 \times 10^{-5} \times (6.4 \times 10^6)^3}{\mu_0 / 4\pi} \\ &= 4 \times 10^2 \times (6.4 \times 10^6)^3 \\ &\quad \left(\frac{\mu_0}{4\pi} = 10^{-7} \right) \\ &= 1.04 \times 10^{23} \text{ A m}^2 \end{aligned}$$

This is close to the value $8 \times 10^{22} \text{ A m}^2$ quoted in geomagnetic texts.

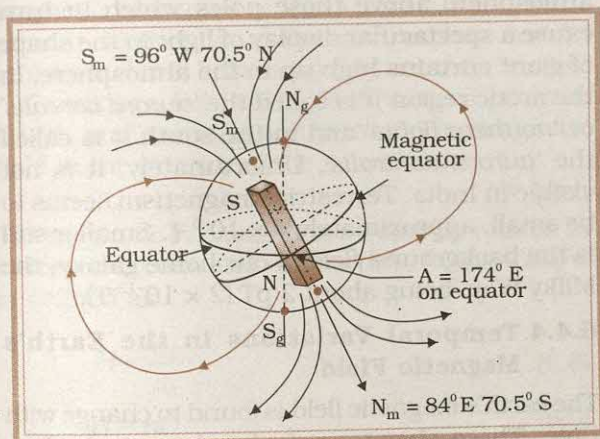


Fig. 6.5 The magnetic field of the earth.

6.4.1 Magnetic Declination and Dip

Consider a point P on the earth's surface, say, Delhi. At this point, the longitude determines the north-south direction. The vertical plane containing the vertical axis and the longitude is called the *geographic meridian*. At P, there also exists the earth's magnetic field \mathbf{B} . The *magnetic meridian* is the vertical plane containing \mathbf{B} and the vertical axis. The angle between the geographic and the magnetic meridian planes is called the *magnetic declination*. The magnetic declination may be determined in the following way: At point P, set up a compass needle in the horizontal plane and free to rotate about the vertical axis. The angle that this needle makes with the geographic north-south direction (N_g-S_g) is the magnetic declination. Such an arrangement is shown in Fig. 6.6(a). The magnetic declination in India is rather small. That of Delhi is $0^\circ 41' \text{ E}$ and of Mumbai is $0^\circ 58' \text{ W}$. This means that the direction of the geographic north is given quite accurately by the compass needle.

There is another quantity of interest, namely the magnetic dip. As we move north from the equator the magnetic field changes direction and dips down. At a point in Northern Canada the magnetic field points straight down into the earth – the magnetic south pole S_m of the earth. We can determine the dip angle at a point on the earth's surface by using a dip circle or dip-meter. We first determine the magnetic declination using the method described above. Thus we know the magnetic meridian. The dip circle is simply a compass needle pivoted about the horizontal axis and free to move in the magnetic meridian. The angle that it makes with the horizontal is called the *dip angle*. This arrangement is shown in Fig. 6.6(b).

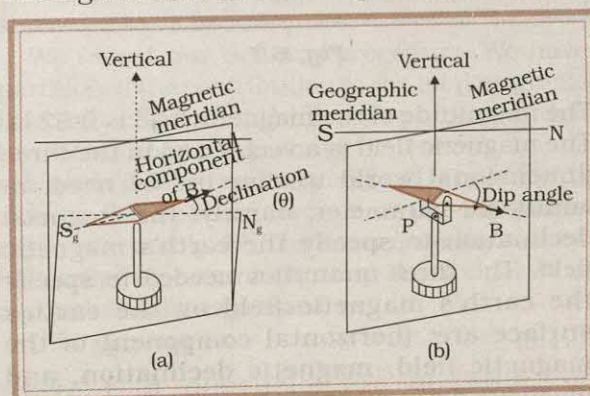


Fig. 6.6 (a) The magnetic meridian and magnetic declination θ is determined by a compass needle free to rotate in the horizontal plane and about the vertical axis. (b) The magnetic dip angle is measured by allowing the compass needle to rotate freely in the vertical plane of the magnetic meridian and about the horizontal axis.

Example 6.5 In the magnetic meridian of a certain place, the horizontal component of the earth's magnetic field is 0.26 G and the dip angle is 60° . What is the magnetic field of the earth in this location?

Answer

It is given that $B_h = 0.26 \text{ G}$. From Fig. 6.7, we have

$$\cos 60^\circ = \frac{B_h}{B}$$

$$B = \frac{B_h}{\cos 60^\circ}$$

$$= \frac{0.26}{(1/2)} = 0.52 \text{ G}$$

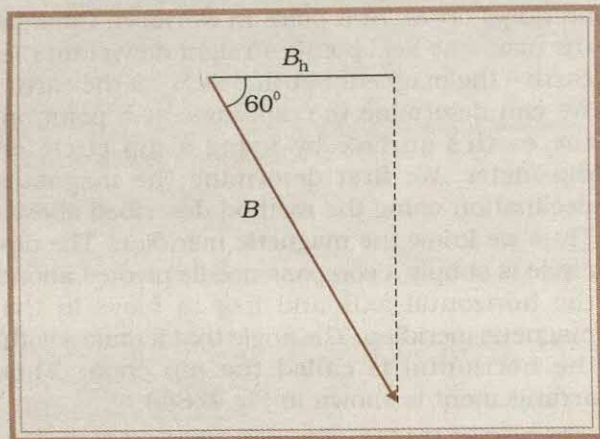


Fig. 6.7

The magnitude of the magnetic field is 0.52 G. The magnetic field is a vector and in the three dimensional world we live in, we need an additional parameter, namely, the magnetic declination to specify the earth's magnetic field. The three quantities needed to specify the earth's magnetic field on the earth's surface are: (horizontal component of the magnetic field, magnetic declination, and magnetic dip).

6.4.2 Origin of the Earth's Magnetic Field

The magnetic field of the earth can be approximated by a giant bar magnet located deep inside the earth as shown in Fig. 6.5. But such an entity, although useful, is entirely fictitious. The earth does have large deposits of iron ore deep inside. But it is highly unlikely that any large solid mass of magnetic material is responsible for the earth's magnetic field. The earth's core is very hot and molten. Circulating ions in the highly conducting liquid region of the earth's core could form current loops and produce a magnetic field. This hypothesis seems probable since our moon which has no molten core, also has no magnetic field. Venus which has a slower rate of rotation, has a weaker magnetic field. On the other hand, Jupiter with a faster rotation rate has a stronger field. However, what is the precise mode of action and the energy needed to sustain such circulating currents? This remains an open question and is a matter of continuing research.

6.4.3 Global Variations in the Earth's Magnetic Field

The dipole approximation (Example 6.4) suggests that the earth's magnetic field falls below a micro-tesla (10^{-6} T) at a distance of five times the earth's radius, i.e., at about 30,000 km. Beyond this, the *solar wind* disturbs the dipole pattern. The solar wind consists of streams of charged particles that emerge continuously from the sun. A long magnetotail stretches out for several thousand earth diameters in a direction away from the sun. The charged particles of the solar wind get trapped near the magnetic poles of the earth. They ionise the atmosphere above these poles which in turn cause a spectacular display of light in the shape of giant curtains high up in the atmosphere. In the arctic region it is called the '*aurora borealis*' or '*northern lights*' and in the south it is called the '*aurora australis*'. Unfortunately, it is not visible in India. Terrestrial magnetism seems to be small, approximately $4 \times 10^{-5} \text{ T}$. Smaller still is the background field of our home galaxy, the Milky Way, being about 2 pT ($2 \times 10^{-12} \text{ T}$).

6.4.4 Temporal Variations in the Earth's Magnetic Field

The earth's magnetic field is found to change with time. These changes may be characterised as short term and long term. In the short term, the magnetic poles of the earth S_m and N_m keep shifting their positions. In a span of 240 years, from 1580 to 1820, the magnetic declination at London has changed by 35° . The magnetic south pole is in the northern Arctic region of Canada. It has been mapped with regularity in the last century. Surprisingly, it has been found to move in a north-westerly direction at a rate of 10 km per year in recent times. The reason for this is not known. Detailed charts of the earth's field are maintained and revised periodically. The magnetic maps of our earth are being drawn with an accuracy no less than the geographical ones.

The changes in the earth's magnetic field over long term or geological time scales are quite interesting. It appears that the earth's field has reversed itself every million year or so! Evidence for this comes from basalt, which contains iron, and is emitted during volcanic activity on the ocean floor. As the basalt cools, it solidifies and provides a picture of the earth's magnetic field direction. The basalt can be dated by other

means and thus a clear picture of the reversal of the earth's magnetic field over geological time scales has emerged. These reversals mean that once in a million years or so, the currents in the earth's core slow down, come to a halt, and then pick up in the opposite direction. The ancient science of geomagnetism continues to surprise and mystify us.

6.5 MAGNETISATION AND MAGNETIC INTENSITY

The earth abounds with a bewildering variety of elements and compounds. In addition, we have been synthesising new alloys, compounds and even elements. One would like to classify the magnetic properties of these substances. In the present section, we define and explain certain terms which will help us to carry out this exercise.

Consider a long solenoid of n turns per unit length and carrying a current i . The magnetic field in the interior of the solenoid is,

$$B_0 = \mu_0 ni \quad (6.7)$$

The *magnetic intensity* \mathbf{H} is a quantity related to currents in coils and conductors. In this case it is defined as

$$H = \frac{B_0}{\mu_0} \quad (6.8)$$

$$H = ni \quad (6.9)$$

The magnetic intensity is a vector with dimensions of $L^{-1}A$. Its SI unit is $A\ m^{-1}$ (ampere metre $^{-1}$).

We next fill the interior of the solenoid with a magnetic material, keeping the current i constant. The total field \mathbf{B} inside will be different

from \mathbf{B}_0 . The additional field will be due to the magnetic material under the influence of \mathbf{B}_0 . Let this magnetic material possess a dipole moment \mathbf{m} . We define a relevant quantity called the *magnetisation* \mathbf{M} which is equal to the magnetic moment per unit volume (V)

$$\mathbf{M} = \frac{\mathbf{m}}{V} \quad (6.10)$$

\mathbf{M} is a vector with dimensions $L^{-1}A$ and units Am^{-1} . Thus \mathbf{M} and \mathbf{H} have the same units. It turns out that the additional magnetic field inside the sample is $\mu_0 \mathbf{M}$. You can check that $\mu_0 \mathbf{M}$ does have the dimensions of magnetic field. Thus, the total field \mathbf{B} is,

$$\begin{aligned} \mathbf{B} &= \mathbf{B}_0 + \mu_0 \mathbf{M} \\ &= \mu_0 (\mathbf{H} + \mathbf{M}) \end{aligned} \quad (6.11)$$

where we have used Eq. 6.8.

We repeat our defining procedure. We have partitioned the contribution to the total magnetic field inside the sample into two parts: one, due to external factors such as the current in the solenoid. This is represented by \mathbf{H} . The other is due to the specific nature of the magnetic material, namely, \mathbf{M} . The latter quantity can be influenced by external factors. This influence is mathematically expressed as

$$\mathbf{M} = \chi \mathbf{H} \quad (6.12)$$

where χ , a dimensionless quantity, is appropriately called the *magnetic susceptibility*. It is a measure of how a magnetic material responds to an external field. Table 6.2 lists χ for some elements. We see that χ is quite small. It is small and positive for materials which are called paramagnetic. It is small and negative

Table 6.2 Magnetic Susceptibility of some Elements at 300 K

Diamagnetic Substance	χ	Paramagnetic Substance	χ
Bismuth	-1.66×10^{-5}	Aluminium	2.3×10^{-5}
Copper	-9.8×10^{-6}	Calcium	1.9×10^{-5}
Diamond	-2.2×10^{-5}	Chromium	2.7×10^{-4}
Gold	-3.6×10^{-5}	Lithium	2.1×10^{-5}
Lead	-1.7×10^{-5}	Magnesium	1.2×10^{-5}
Mercury	-2.9×10^{-5}	Niobium	2.6×10^{-5}
Nitrogen (STP)	-5.0×10^{-9}	Oxygen (STP)	2.1×10^{-6}
Silver	-2.6×10^{-5}	Platinum	2.9×10^{-4}
Silicon	-4.2×10^{-6}	Tungsten	6.8×10^{-5}

for materials which are termed diamagnetic. In the latter case \mathbf{M} and \mathbf{H} are opposite in direction. From Eqs. (6.11) and (6.12) we obtain,

$$B = \mu_0 (1 + \chi)H$$

$$= \mu_0 \mu_r H \quad (6.13)$$

$$= \mu H \quad (6.14)$$

Where $\mu_r = 1 + \chi$, is a dimensionless quantity called the *relative magnetic permeability* of the substance. It is the analog of the dielectric constant in electrostatics. The *magnetic permeability* of the substance is μ and it has the same dimensions and units as μ_0 ; $\mu = \mu_0 \mu_r = \mu_0 (1 + \chi)$.

The three quantities χ , μ_r and μ are inter-related and only one of them is independent. Given one, the other two may be easily determined.

Example 6.6 Obtain the earth's magnetisation. Assume that the earth's field can be approximated by a giant bar magnet of magnetic moment $8.0 \times 10^{22} \text{ A m}^2$. The earth's radius is 6400 km.

Answer The earth's radius $R = 6400 \text{ km}$.

$$R = 6.4 \times 10^6 \text{ m}$$

Magnetisation is the magnetic moment per unit volume. Hence,

$$M = \frac{m}{\frac{4\pi}{3} R^3}$$

$$= \frac{8.0 \times 10^{22} \times 3}{4 \times \pi \times (6.4 \times 10^6)^3}$$

$$= \frac{24.0 \times 10^4}{4 \times \pi \times 262.1} = 72.9 \text{ A m}^{-1}$$

Example 6.7 A solenoid of 500 turns/m is carrying a current of 3 A. Its core is made of iron which has a relative permeability of 5000. Determine the magnitudes of the magnetic intensity, magnetization and the magnetic field inside the core.

Answer From Eq. (6.9) the magnetic intensity is,

$$H = n i$$

$$= 500 \text{ m}^{-1} \times 3 \text{ A}$$

$$= 1500 \text{ A m}^{-1}$$

It is given that $\mu_r = 5000$. From Eqs. (6.13) and (6.14),

$$\mu_r = 1 + \chi, \text{ i.e., } \chi = 4999 \approx 5000$$

$$\mu = 5000 \mu_0$$

Hence, the magnetisation is, from Eq. (6.12),

$$M = \chi H$$

$$= 7.5 \times 10^6 \text{ A m}^{-1}$$

The magnetic field is, from Eq. (6.14),

$$B = 5000 \mu_0 H$$

$$= 5000 \times 4\pi \times 10^{-7} \times 1500$$

$$= 9.4 \text{ T.}$$

6.6 MAGNETIC PROPERTIES OF MATERIALS

The discussion in the previous section helps us to classify materials as diamagnetic, paramagnetic or ferromagnetic. In terms of the susceptibility χ , a material is diamagnetic if χ is negative, para- if χ is positive and small, and ferro- if χ is large and positive. More concretely

Diamagnetic	$-1 \leq \chi < 0$	$0 \leq \mu_r < 1$	$\mu < \mu_0$
Paramagnetic	$0 < \chi < \varepsilon$	$1 < \mu_r < 1 + \varepsilon$	$\mu > \mu_0$
Ferromagnetic	$\chi \gg 1$	$\mu_r \gg 1$	$\mu \gg \mu_0$

Here ε is a small positive number introduced to quantify paramagnetic materials. A glance at Table 6.2 gives one a better feeling for these materials. Next we describe these materials in some detail.

6.6.1 Diamagnetism

The individual atoms (or ions or molecules) of a diamagnetic material do not possess a permanent dipole moment of their own. The application of an external magnetic field \mathbf{B}_0 induces in each atom, a small dipole moment proportional to B_0 , but pointing in the opposite direction. Figure 6.8(a) shows a bar of diamagnetic material placed in an external field. The field lines are repelled or expelled and the field inside the material is reduced. In most cases, as is evident from Table 6.2, this reduction is slight, being one part in 10^5 . When placed in a non-uniform magnetic field, the bar will tend to move from high field to low.

Some diamagnetic materials are bismuth, copper, lead, silicon, nitrogen (at STP), water and sodium chloride. The most exotic diamagnetic

materials are Type-I superconductors. These are metals, cooled to very low temperatures which exhibit both *perfect conductivity* and *perfect diamagnetism*. Here the field lines are completely expelled! $\chi = -1$ and $\mu_r = 0$. A superconductor repels a magnet and (by Newton's III law) is repelled by the magnet. The phenomenon of perfect diamagnetism in superconductors is called the *Meissner effect*, so named after its discoverer. It can be gainfully exploited in a variety of situations, e.g., for running magnetically levitated superfast trains.

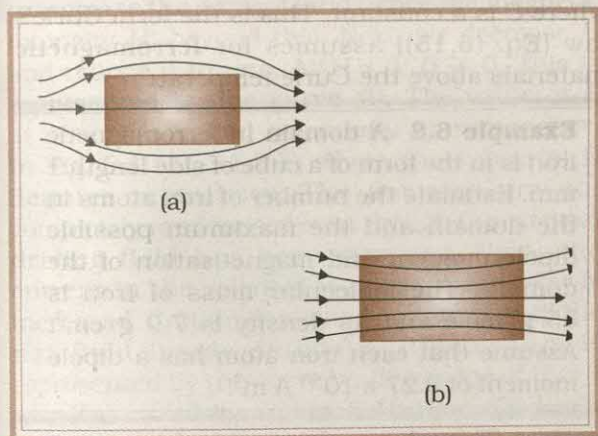


Fig. 6.8 Behaviour of field lines due to an external magnetic field near a (a) diamagnetic (b) paramagnetic substance.

6.6.2 Paramagnetism

The individual atoms (or ions or molecules) of a paramagnetic material possess a permanent dipole moment of their own. On account of the ceaseless random thermal motion of the atoms, no net magnetization is seen. In the presence of an external field \mathbf{B}_0 which is strong enough, and at low temperatures, the individual atomic dipole moments can be made to align and point in the same direction as \mathbf{B}_0 . Figure 6.8(b) shows a bar of paramagnetic material placed in an external field. The field lines get concentrated inside the material, and the field inside is enhanced. In most cases, as is evident from Table 6.2, this enhancement is slight, being one part in 10^5 . When placed in a non-uniform magnetic field, the bar will tend to move from low field to high.

Some paramagnetic materials are aluminium, sodium, calcium, oxygen (at STP)

and copper chloride. Experimentally, one finds that the magnetisation of a paramagnetic material is directly proportional to the applied field and inversely proportional to the absolute temperature T .

$$M = C \frac{B_0}{T} \quad [6.15(a)]$$

or equivalently, using Eqs. (6.8) and (6.12)

$$\chi = C \frac{\mu_0}{T} \quad [6.15(b)]$$

This is known as Curie's law, after its discoverer Pierre Curie (1859-1906). The constant C is called Curie's constant. Thus, for a paramagnetic material both χ and μ_r depend not only on the material, but also (in a simple fashion) on the sample temperature. At very high fields or at very low temperatures, the magnetisation approaches its maximum value when all atomic dipole moments are aligned. This is called the saturation magnetisation value M_s . Beyond this, Curie's law [Eq. (6.15)] is no longer valid.

6.6.3 Ferromagnetism

The individual atoms (or ions or molecules) in a ferromagnetic material possess a dipole moment, as in a paramagnetic material. However, they interact with one another in such a way that they spontaneously align themselves in a common direction over a macroscopic volume called *domain*. The explanation for this co-operative effect requires quantum mechanics and is beyond the scope of this textbook. Each domain has a net magnetisation. Typical domain size is 1 mm and the domain contains about 10^{11} atoms. In the first instance, the magnetisation varies randomly from domain to domain and there is no bulk magnetisation. This is shown in Fig. 6.9(a). When we apply an external magnetic field \mathbf{B}_0 , the domains orient themselves in the direction of \mathbf{B}_0 and simultaneously the domains oriented in the direction of \mathbf{B}_0 grow in size. The existence of domains and their motion in \mathbf{B}_0 are not speculations. One may observe these under a microscope after sprinkling a liquid suspension of powdered ferromagnetic substance on the sample. The motion of this suspension can be observed. Figure 6.9(b) shows the situation when the domains have aligned and amalgamated to form a single 'giant' domain.

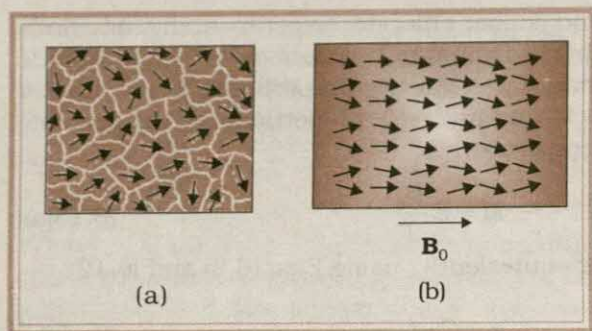


Fig. 6.9 (a) Randomly oriented domains; (b) Aligned domains.

Thus in a ferromagnetic material the field lines are highly concentrated. In a non-uniform magnetic field the sample tends to move towards the region of high field. We may wonder as to what happens when the external field is removed? In some ferromagnetic materials, the magnetisation persists. Such materials are called 'hard' magnetic materials or hard ferromagnets. Alnico, an alloy of iron, aluminium, nickel, cobalt and copper, is one such material. The naturally occurring lodestone is another. Such materials form permanent magnets to be used among other things as a compass needle. On the other hand, there is a class of ferromagnetic materials in which the magnetisation disappears on removal of the external field. Soft iron is one such material. Appropriately enough, such materials are called soft ferromagnetic materials. There are a number of elements which are ferromagnetic: iron, cobalt, nickel, gadolinium, etc. The relative magnetic permeability is >1000 !

The ferromagnetic property depends on temperature. At high enough temperatures, a ferromagnet becomes a paramagnet. The domain structure disintegrates with temperature. This disappearance of magnetisation with temperature is gradual. It is a phase transition reminding us of the melting of a solid crystal. The temperature of transition from ferromagnetism to paramagnetism is called the Curie temperature T_c . Table 6.3 lists the Curie temperature of certain ferromagnets.

The susceptibility *above* the Curie temperature, i.e., in the paramagnetic phase is described by,

$$\chi = \frac{C}{T - T_c} \quad (T > T_c)$$

Table 6.3 Curie Temperature T_c of some Ferromagnetic Materials

Material	T_c (K)
Cobalt	1394
Iron	1043
Fe_2O_3	893
Nickel	631
Gadolinium	317

where C is a constant. This is the form Curie's law [Eq. (6.15)] assumes for ferromagnetic materials above the Curie temperature.

Example 6.8 A domain in ferromagnetic iron is in the form of a cube of side length 1 mm. Estimate the number of iron atoms in the domain and the maximum possible dipole moment and magnetisation of the domain. The molecular mass of iron is 55 g/mole and its density is 7.9 g/cm³. Assume that each iron atom has a dipole moment of 9.27×10^{-24} A m².

Answer The volume of the cubic domain is

$$V = (10^{-6})^3 = 10^{-18} \text{ m}^3 = 10^{-12} \text{ cm}^3$$

Its mass is volume \times density

$$= 7.9 \times 10^{-12} \text{ g}$$

It is given that an Avogadro number (6.023×10^{23}) of iron atoms has a mass of 55 g. Hence, the number of atoms in the domain is

$$N = \frac{7.9 \times 10^{-12} \times 6.023 \times 10^{23}}{55} \\ = 8.65 \times 10^{10} \text{ atoms}$$

The maximum possible dipole moment m_{max} is achieved for the (unrealistic) case when all the atomic moments are perfectly aligned. Thus,

$$m_{\text{max}} = (8.65 \times 10^{10}) \times (9.27 \times 10^{-24}) \\ = 8.0 \times 10^{-13} \text{ A m}^2$$

The consequent magnetisation is

$$M_{\text{max}} = \frac{m_{\text{max}}}{\text{Domain volume}} \\ = \frac{8.0 \times 10^{-13}}{10^{-18}} \\ = 8.0 \times 10^5 \text{ A m}^{-1}.$$

The relation between \mathbf{B} and \mathbf{H} in ferromagnetic materials is complex. It is often not linear and it depends on the magnetic history of the sample. Fig. 6.10 depicts the behaviour of the material as we take it through one cycle of magnetisation. Let the material be unmagnetised initially. We place it in a solenoid and increase the current through the solenoid. The magnetic field B in the material rises and saturates as depicted in the curve Oa. This behaviour represents the alignment and merger of domains until no further enhancement is possible. It is pointless to increase the current (and hence the magnetic intensity H) beyond this. Next, we decrease H and reduce it to zero. At $H = 0$, $B \neq 0$. This is represented by the curve ab. The value of B at $H = 0$ is called *retentivity* or *remanence*. In Fig. 6.10 $B_R \sim 1.2$ T, where the subscript R denotes retentivity. The domains are not completely randomised even though the external driving field has been removed. Next, the current in the solenoid is reversed and slowly increased. Certain domains are flipped until the net field inside stands nullified. This is represented by the curve bc. The value of H at c is called *coercivity*. In Fig. 6.10 $H_c \sim -90 \text{ A m}^{-1}$. As the reversed current is increased in magnitude, we once again obtain saturation. The (curve cd) depicts this. The saturated magnetic field $B_s \sim 1.5$ T. Next, the current is reduced (curve de) and reversed (curve ea). The cycle repeats itself. We make the following observations:

- (i) The curve Oa does not retrace itself as H is reduced. For a given value of H , B is not unique, but depends on the previous history of the sample. This phenomenon is called *hysteresis*. The word hysteresis means 'lagging behind' (and not 'history').
- (ii) In Fig. 6.10 when $B \approx 1.5$ T, we obtain $B/\mu_0 \approx 1.25 \times 10^6 \text{ A m}^{-1}$. However, H at this value of B is merely 200 A m^{-1} . This implies that the effective relative magnetic permeability $\mu_r \approx 10000$! Such large values are characteristic of ferromagnetic materials.
- (iii) No segment, Oa, ab, etc. of the curve is linear. B is not proportional to H over any appreciable range.
- (iv) Note that $BH = B^2/(\mu_0 \mu_r)$ has the dimensions of energy per unit volume. The

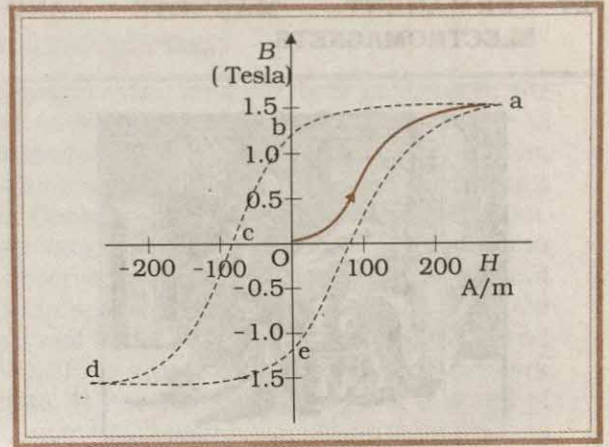


Fig. 6.10 The magnetic hysteresis loop is the B - H curve for ferromagnetic materials.

area within the B - H loop of Fig. 6.10 represents the energy dissipated per unit volume in the material. The source of energy is the chemical battery which drives the current in the solenoid. The 'sink' is the Joule heating ($i^2 R$) in the solenoid wire and the dissipative hysteretic heat loss in the magnetic material. A 'hard' ferromagnetic material has a wide hysteresis loop as shown in Fig. 6.11(a). It has large coercivity and is useful in designing permanent magnets. A 'soft' ferromagnet has a narrow hysteresis loop as depicted in Fig. 6.11(b). The energy loss is minimal and at the same time the value of μ_r enhances the magnetic field. Such 'soft' magnetic materials are useful as cores in solenoids and transformers where field reversals are frequent.

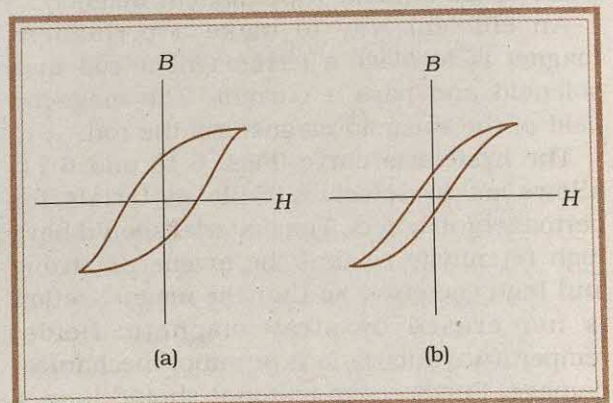


Fig. 6.11 The magnetic hysteresis loop for (a) hard; (b) soft ferromagnetic material.

6.7 PERMANENT MAGNETS AND ELECTROMAGNETS



Fig. 6.12 A blacksmith forging a permanent magnet by striking a red-hot rod of iron kept in the north-south direction with a hammer. The sketch is recreated from an illustration in *De Magnete*, a work published in 1600 and authored by William Gilbert, the court physician to Queen Elizabeth of England.

Substances which at room temperature retain their ferromagnetic property for a long period of time are called permanent magnets. Permanent magnets can be made in a variety of ways. One can hold an iron rod in the north-south direction and hammer it repeatedly. The method is illustrated in Fig. 6.12. The illustration is from a 400 year old book to emphasise that the making of permanent magnets is an old 'art'. One can also hold a steel rod and stroke it with one end of a bar magnet a large number of times, always in the same sense to make a permanent magnet.

An efficient way to make a permanent magnet is to place a ferromagnetic rod in a solenoid and pass a current. The magnetic field of the solenoid magnetises the rod.

The hysteresis curve (Figs. 6.10 and 6.11) allows us to select suitable materials for permanent magnets. The material should have high retentivity so that the magnet is strong and high coercivity so that the magnetisation is not erased by stray magnetic fields, temperature fluctuations or minor mechanical damage. Further, the material should have a

high permeability. Steel is one favoured choice. It has a slightly smaller retentivity than soft iron but this is outweighed by the much smaller coercivity of soft iron. Other suitable material for permanent magnets are alnico (an alloy of iron, aluminium, nickel, cobalt, and copper), cobalt steel and ticonal.

Electromagnets are made of ferromagnetic materials which have high permeability and low retentivity. Soft iron is a suitable material for electromagnets. On placing a soft iron rod in a solenoid and passing a current, we increase the magnetism of the solenoid by a thousand fold. When we switch off the solenoidal current, the magnetism is effectively switched off since the soft iron core has a low retentivity. The arrangement is shown in Fig. 6.13.

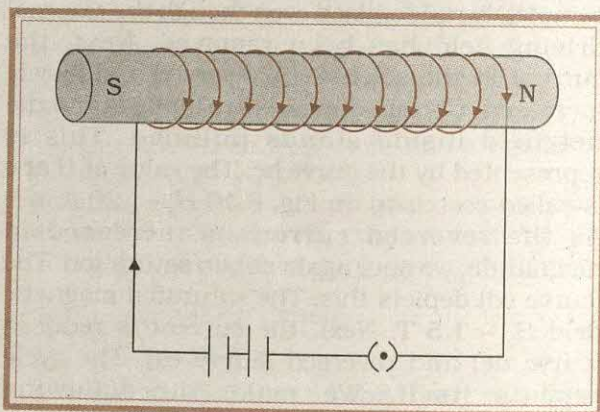


Fig. 6.13 A soft iron core in solenoid acts as an electromagnet.

In certain applications, the material goes through an ac cycle of magnetisation for a long period. This is the case in transformer cores and telephone diaphragms. The hysteresis curve of such materials must be narrow. The energy dissipated and the heating will consequently be small. The material must have a low resistivity to lower eddy current losses*.

Electromagnets are used in electric bells, loudspeakers and telephone diaphragms. Giant electromagnets are used in cranes to lift machinery, and bulk quantities of iron and steel.

* We shall study about eddy currents in Chapter 7.

MAPPING INDIA'S MAGNETIC FIELD

Because of its practical application in prospecting, communication, and navigation, the magnetic field of the earth is mapped by most nations with an accuracy comparable to geographical mapping. In India over a dozen observatories exist, extending from Trivandrum (now Thiruvananthapuram) in the south to Gulmarg in the north. These observatories work under the aegis of the Indian Institute of Geomagnetism (IIG), in Colaba, Mumbai. The IIG grew out of the Colaba and Alibag observatories and was formally established in 1971. The IIG monitors (via its nationwide observatories), the geomagnetic fields and fluctuations on land, and under the ocean and in space. Its services are used by the Oil and Natural Gas Commission (ONGC), the National Institute of Oceanography (NIO) and the Indian Space Research Organisation (ISRO). It is a part of the world-wide network which ceaselessly updates the geomagnetic data. It has participated in over a dozen of India's expeditions to the Antarctic region. Geomagnetic research is a vibrant and exciting scientific field.

SUMMARY

1. The science of magnetism is old. It has been known since ancient times that magnetic materials tend to point in the north-south direction; like magnetic poles repel and unlike ones attract; and cutting a bar magnet in two leads to two smaller magnets. Magnetic poles cannot be isolated.
2. When a bar magnet of dipole moment \mathbf{m} is placed in a uniform magnetic field \mathbf{B} ,
 - (a) the force on it is zero,
 - (b) the torque on it is $\mathbf{m} \times \mathbf{B}$.
 - (c) its potential energy is $-\mathbf{m} \cdot \mathbf{B}$, where we choose the zero of energy at the orientation when \mathbf{m} is perpendicular to \mathbf{B} .
3. Consider a bar magnet of size l and magnetic moment \mathbf{m} . At a distance r from its mid-point, where $r \gg l$, the magnetic field \mathbf{B} due to this bar is

$$\mathbf{B} = \frac{\mu_0 \mathbf{m}}{2\pi r^3} \quad (\text{along axis})$$

$$= -\frac{\mu_0 \mathbf{m}}{4\pi r^3} \quad (\text{along equator})$$

4. Gauss's law for magnetism states that the net magnetic flux through any closed surface is zero

$$\phi_B = \sum_{\substack{\text{all area} \\ \text{elements} \Delta \mathbf{S}}} \mathbf{B} \cdot \Delta \mathbf{S} = 0$$

5. The earth's magnetic field may be approximated by a dipole with moment $8.0 \times 10^{22} \text{ A m}^{-1}$. This dipole is aligned making a small angle with the rotation axis of the earth. Its *magnetic north pole* N_m is near the *geographic south pole* S_g and its *magnetic south pole* S_m is near the *geographic north pole* N_g . The magnitude of the field on the earth's surface $\approx 4 \times 10^{-5} \text{ T}$.
6. Three quantities are needed to specify the magnetic field of the earth on its surface — the horizontal component, the magnetic declination, and the magnetic dip. Please refer to the text for precise definitions of declination and dip.

7. Consider a material placed in an external magnetic field \mathbf{B}_0 . The magnetic intensity is defined as,

$$\mathbf{H} = \frac{\mathbf{B}_0}{\mu_0}$$

The magnetisation \mathbf{M} of the material is its dipole moment per unit volume. The magnetic field \mathbf{B} in the material is,

$$\mathbf{B} = \mu_0(\mathbf{H} + \mathbf{M})$$

8. For a linear material $\mathbf{M} = \chi \mathbf{H}$. So that $\mathbf{B} = \mu \mathbf{H}$ and χ is called the magnetic susceptibility of the material. The three quantities, χ , the relative magnetic permeability μ_r , and the magnetic permeability μ are related as follows:

$$\mu = \mu_0 \mu_r$$

$$\mu_r = 1 + \chi$$

9. Magnetic materials are broadly classified as: diamagnetic, paramagnetic, and ferromagnetic. For diamagnetic materials χ is negative and generally very small, for paramagnetic χ is positive and small. Ferromagnetic materials have large χ and are similarly characterised by non-linear relation between \mathbf{B} and \mathbf{H} . They show the property of hysteresis as explained in the text.

Physical quantity	Symbol	Nature	Dimensions	Units	Remarks
Permeability of free space	μ_0	Scalar	$[\text{MLT}^{-2}\text{A}^{-2}]$	T m A^{-1} N A^{-2}	$\mu_0/4\pi = 10^{-7}$
Magnetic Field, Magnetic Induction, Magnetic Flux Density	\mathbf{B}	Vector	$[\text{MT}^{-2}\text{A}^{-1}]$	T (tesla)	$10^4 \text{ G (gauss)} = 1 \text{ T}$
Magnetic Moment	\mathbf{m}	Vector	$[\text{L}^2\text{A}]$	A m^2	
Magnetic Flux	ϕ_B	Scalar	$[\text{ML}^2\text{T}^{-2}\text{A}^{-1}]$	W (weber)	$W = \text{T m}^2$
Magnetisation	\mathbf{M}	Vector	$[\text{L}^{-1}\text{A}]$	A m^{-1}	$\frac{\text{Magnetic moment}}{\text{Volume}}$
Magnetic Intensity Magnetic Field Strength	\mathbf{H}	Vector	$[\text{L}^{-1}\text{A}]$	A m^{-1}	$\mathbf{B} = \mu_0(\mathbf{H} + \mathbf{M})$
Magnetic Susceptibility	χ_m	Scalar	-	-	$\mathbf{M} = \chi_m \mathbf{H}$
Relative magnetic permeability	μ_r	Scalar	-	-	$\mathbf{B} = \mu_0 \mu_r \mathbf{H}$
Magnetic permeability	μ	Scalar	$[\text{MLT}^{-2}\text{A}^{-2}]$	T m A^{-1} N A^{-2}	$\mu = \mu_0 \mu_r$ $\mathbf{B} = \mu \mathbf{H}$

POINTS TO PONDER

1. A satisfactory understanding of magnetic phenomena in terms of moving charges/currents was arrived at after 1800. But *technological exploitation* of the directional properties of magnets predates this *scientific understanding* by two thousand years. Thus, scientific understanding is not a necessary condition for engineering applications. Ideally, science and engineering go hand-in-hand, one leading and assisting the other in tandem.
2. Magnetic monopoles do not exist. If you slice a magnet in half, you get two smaller magnets. On the other hand, isolated positive and negative charges exist. There exists a smallest unit of charge, e.g., the electronic charge with value $|e| = 1.6 \times 10^{-19}$ C. All other charges are integral multiples of this smallest unit charge. In other words, charge is quantised. We do not know why magnetic monopoles do not exist or why electric charge is quantised. The physicist Paul Dirac speculated that the two unexplained facts are inter-related. He demonstrated that if a magnetic monopole exists then charge will be quantised.
3. A consequence of the fact that magnetic monopoles do not exist is that the magnetic field lines are continuous and form closed loops. In contrast, the electrostatic lines of force begin on a positive charge and terminate on the negative charge (or fade out at infinity).
4. The earth's magnetic field is not due to a huge bar magnet inside it. The earth's core is hot and molten. Perhaps convective currents in this core are responsible for the earth's magnetic field. As to what 'dynamo' effect sustains this current, and why the earth's field reverses polarity every million years or so, we do not know.
5. A miniscule difference in the value of χ , the magnetic susceptibility, yields radically different behaviour: diamagnetic versus paramagnetic. For diamagnetic materials $\chi \simeq -10^{-5}$ whereas $\chi \simeq +10^{-5}$ for paramagnetic materials.
6. There exists a *perfect diamagnet*, namely, the type-I superconductor. This is a metal at very low temperatures. In this case $\chi = -1$, $\mu_r = 0$, $\mu = 0$. The external magnetic field is totally expelled. Interestingly, this material is also a *perfect conductor*. However, there exists no classical theory which ties these two properties together. A quantum-mechanical theory by Bardeen, Cooper, and Schrieffer (BCS theory) explains these effects. The BCS theory was proposed in 1957 and was eventually recognised by a Nobel Prize in Physics in 1970.
7. The phenomenon of magnetic hysteresis is reminiscent of similar behaviour concerning the elastic properties of materials. Strain may not be proportional to stress; here H and B (or M) are not linearly related. The stress-strain curve exhibits hysteresis and the area enclosed by it represents the energy dissipated per unit volume. A similar interpretation can be given to the B - H magnetic hysteresis curve.
8. Diamagnetism is universal. It is present in all materials. But it is weak and hard to detect if the substance is para-or ferromagnetic.
9. We have classified materials as diamagnetic, paramagnetic, and ferromagnetic. However, there exist additional types of magnetic material such as ferrimagnetic, anti-ferromagnetic, spin glass, etc. with properties which are exotic and mysterious.

EXERCISES

6.1 Figure 6.14 shows a small magnetised needle A placed at a point O. The arrow shows the direction of its magnetic moment. The other arrows show different positions (and orientations) of the magnetic moment) of another identical magnetised needle B.

- In which configurations is the system not in equilibrium?
- In which configuration is the system in (i) stable, and (ii) unstable equilibrium?
- Which configuration corresponds to the lowest potential energy among all the configurations shown?

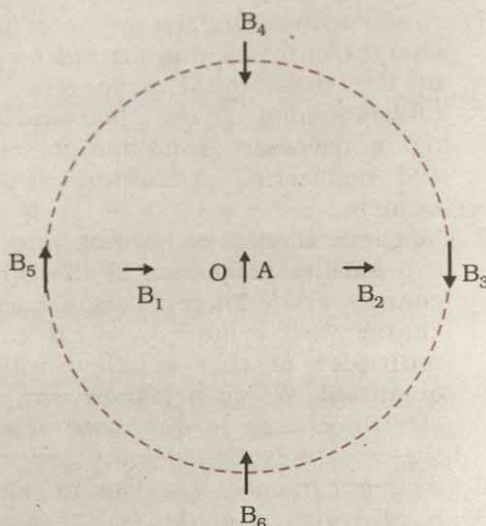


Fig. 6.14

6.2 Answer the following questions:

- What happens if a bar magnet is cut into two pieces (i) transverse to its length (ii) along its length?
- What happens if an iron bar magnet is melted? Does it retain its magnetism?
- A magnetised needle in a uniform magnetic field experiences a torque but no net force. An iron nail near a bar magnet, however, experiences a force of attraction in addition to a torque. Why?
- Must every magnetic field configuration have a north pole and a south pole? What about the field due to a toroid?
- Can you think of a magnetic field configuration with three poles?
- Two identical looking iron bars A and B are given, one of which is definitely known to be magnetised. (We do not know which one). How would one ascertain whether or not both are magnetised? If only one is magnetised, how does one ascertain which one? [Use nothing else but the two bars A and B].

6.3 Answer the following questions regarding earth's magnetism:

- A vector needs three quantities for its specification. Name the three independent quantities conventionally used to specify the earth's magnetic field.
- The angle of dip at a location in southern India is about 18° . Would you expect a greater or smaller dip angle in Britain?
- If you made a map of magnetic field lines at Melbourne in Australia, would the lines seem to go into the ground or come out of the ground?
- In which direction would a compass free to move in the vertical plane point to, if located right on the geomagnetic north or south pole?
- The earth's field, it is claimed, roughly approximates the field due to a dipole of magnetic moment $8 \times 10^{22} \text{ JT}^{-1}$ located at its centre. Check the order of magnitude of this number in some way.
- Geologists claim that besides the main magnetic N-S poles, there are several local poles on the earth's surface oriented in different directions. How is such a thing possible at all?

6.4 Answer the following questions:

- The earth's magnetic field varies from point to point in space. Does it also change with time? If so, on what time scale does it change appreciably?
- The earth's core is known to contain iron. Yet geologists do not regard this as a source of the earth's magnetism. Why?
- The charged currents in the outer conducting regions of the earth's core are thought to be responsible for earth's magnetism. What might be the 'battery' (i.e. the source of energy) to sustain these currents?
- The earth may have even reversed the direction of its field several times during its history of 4 to 5 billion years. How can geologists know about the earth's field in such distant past?
- The earth's field departs from its dipole shape substantially at large distances (greater than about 30,000 km). What agencies may be responsible for this distortion?
- Interstellar space has an extremely weak magnetic field of the order of 10^{-12} T. Can such a weak field be of any significant consequence? Explain.

[**Note:** Exercise 6.4 is meant mainly to arouse your curiosity. Answers to some questions above are tentative or unknown. Brief answers wherever possible are given at the end. For details, you should consult a good text on geomagnetism.]

- 6.5** Many of the figures given in Fig. 6.15 show magnetic field lines *wrongly* (thick lines in the figure). Point out what is wrong with them. Some of them may describe electrostatic field lines correctly. Point out which ones.

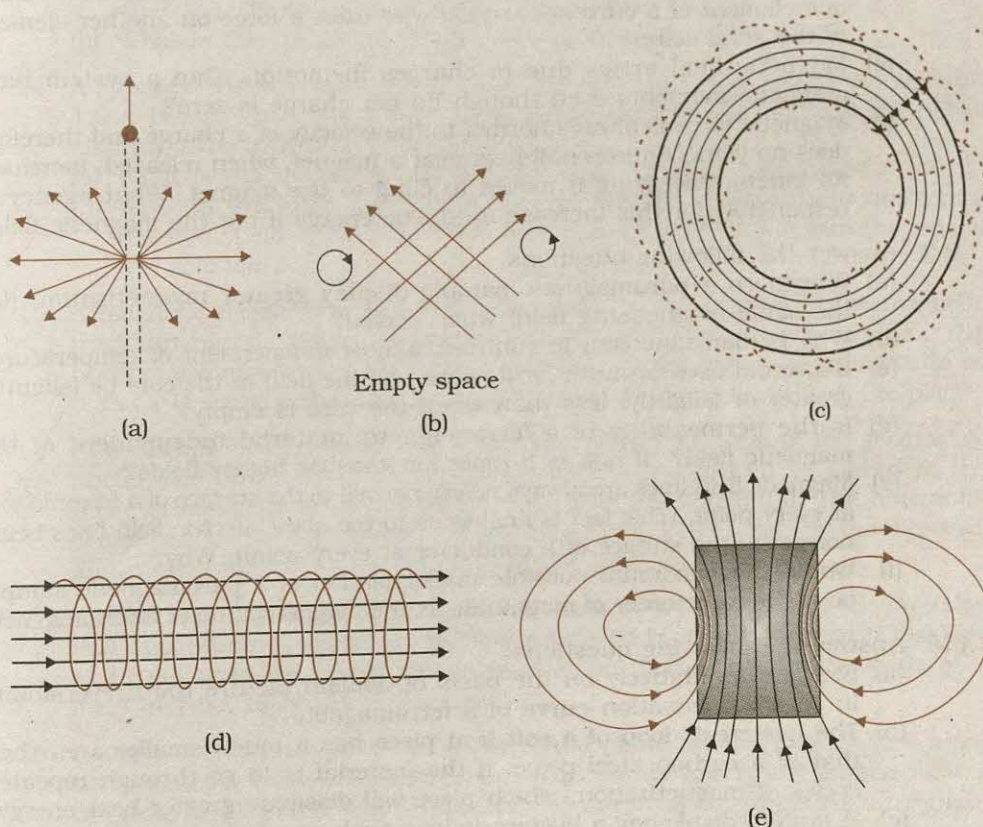


Fig. 6.15 (a), (b), (c), (d), (e)

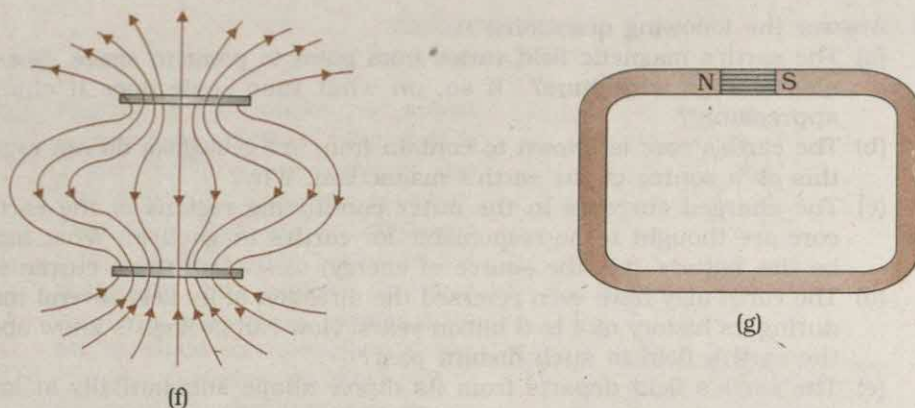


Fig. 6.15 (f) and (g)

6.6 Answer the following questions carefully:

- Magnetic field lines show the direction (at every point) which a small magnetised needle takes up (at that point). Do the magnetic field lines also represent the 'lines of force' on a moving charged particle at every point?
- Magnetic field lines can be entirely confined within the core of a toroid, but not within a straight solenoid. Why?
- If magnetic monopoles existed, how would Gauss's law of magnetism be modified?
- Does a bar magnet exert a torque on itself due to its own field? Does one element of a current-carrying wire exert a force on another element of the *same wire*?
- Magnetic field arises due to charges in motion. Can a system have magnetic moments even though its net charge is zero?
- Magnetic force is always normal to the velocity of a charge and therefore does no work. An iron nail held near a magnet, when released, increases its kinetic energy as it moves to cling to the magnet. What agency is responsible for this increase in kinetic energy if not the magnetic field?

6.7 Answer the following questions:

- Why does a paramagnetic sample display greater magnetisation (for the same magnetising field) when cooled?
- Why is diamagnetism, in contrast, almost independent of temperature?
- If a toroid uses bismuth for its core, will the field in the core be (slightly) greater or (slightly) less than when the core is empty?
- Is the permeability of a ferromagnetic material independent of the magnetic field? If not, is it more for lower or higher fields?
- Magnetic field lines are always nearly normal to the surface of a ferromagnet at every point. (This fact is analogous to the static electric field lines being normal to the surface of a conductor at every point). Why?
- Would the maximum possible magnetisation of a paramagnetic sample be of the same order of magnitude as the magnetisation of a ferromagnet?

6.8 Answer the following questions:

- Explain qualitatively on the basis of domain picture the irreversibility in the magnetisation curve of a ferromagnet.
- The hysteresis loop of a soft iron piece has a much smaller area than that of a carbon steel piece. If the material is to go through repeated cycles of magnetization, which piece will dissipate greater heat energy?
- 'A system displaying a hysteresis loop such as a ferromagnet is a device for storing memory?' Explain the meaning of this statement.

- (d) What kind of ferromagnetic material is used for coating magnetic tapes in a cassette player, or for building 'memory stores' in a modern computer?
- (e) A certain region of space is to be shielded from magnetic fields. Suggest a method.
- 6.9** A short bar magnet placed with its axis at 30° with a uniform external magnetic field of 0.25 T experiences a torque of magnitude equal to 4.5×10^{-2} J. What is the magnitude of magnetic moment of the magnet?
- 6.10** A short bar magnet of magnetic moment $m = 0.32 \text{ JT}^{-1}$ is placed in a uniform external magnetic field of 0.15 T. If the bar is free to rotate in the plane of the field, which orientations would correspond to its (i) stable and (ii) unstable equilibrium? What is the potential energy of the magnet in each case?
- 6.11** A closely wound solenoid of 800 turns and area of cross-section $2.5 \times 10^{-4} \text{ m}^2$ carries a current of 3.0 A. Explain the sense in which the solenoid acts like a bar magnet. What is its associated magnetic moment?
- 6.12** If the solenoid in Exercise 6.11 is free to turn about the vertical direction and a uniform horizontal magnetic field of 0.25 T is applied, what is the magnitude of the torque on the solenoid when its axis makes an angle of 30° with the direction of the applied field?
- 6.13** A bar magnet of magnetic moment 1.5 JT^{-1} lies aligned with the direction of a uniform magnetic field of 0.22 T.
- (a) What is the amount of work required by an external torque to turn the magnet so as to align its magnetic moment, (i) normal to the field direction, (ii) opposite to the field direction?
- (b) What is the torque on the magnet in cases (i) and (ii)?
- 6.14** A closely wound solenoid of 2000 turns and area of cross-section $1.6 \times 10^{-4} \text{ m}^2$, carrying a current of 4.0 A, is suspended through its centre allowing it to turn in a horizontal plane.
- (a) What is the magnetic moment associated with the solenoid?
- (b) What is the force and torque on the solenoid if a uniform horizontal magnetic field of $7.5 \times 10^{-2} \text{ T}$ is set up at an angle of 30° with the axis of the solenoid?
- 6.15** A magnetic needle free to rotate in a vertical plane parallel to the magnetic meridian has its north tip pointing down at 22° with the horizontal. The horizontal component of the earth's magnetic field at the place is known to be 0.35 G. Determine the magnitude of the earth's magnetic field at the place.
- 6.16** At a certain location in Africa, a compass points 12° west of the geographic north. The north tip of the magnetic needle of a dip circle placed in the plane of magnetic meridian points 60° above the horizontal. The horizontal component of the earth's field is measured to be 0.16 G. Specify the direction and magnitude of the earth's field at the location.
- 6.17** A short bar magnet has a magnetic moment of 0.48 JT^{-1} . Give the direction and magnitude of the magnetic field produced by the magnet at a distance of 10 cm from the centre of the magnet on (i) the axis (ii) the equatorial lines (normal bisector) of the magnet.
- 6.18** A short bar magnet placed in a horizontal plane has its axis aligned along the magnetic north-south direction. Null points are found on the axis of the magnet at 14 cm from the centre of the magnet. The earth's magnetic field at the place is 0.36 G and the angle of dip is zero. What is the total magnetic field on the normal bisector of the magnet at the same distance

as the null-points (i.e., 14 cm) from the centre of the magnet?

- 6.19** If the bar magnet in Exercise 6.18 is turned around by 180° , where will the new null-points be located?
- 6.20** A short bar magnet of magnetic moment $5.25 \times 10^{-2} \text{ J T}^{-1}$ is placed with its axis perpendicular to the earth's field direction. At what distance from the centre of the magnet, the resultant field is inclined at 45° with the earth's field on (a) its normal bisector and (b) its axis. Magnitude of the earth's field at the place is given to be 0.42 G. Ignore the length of the magnet in comparison to the distances involved.
- 6.21** A circular coil of 16 turns and radius 10 cm carrying a current of 0.75 A rests with its plane normal to an external field of magnitude $5.0 \times 10^{-2} \text{ T}$. The coil is free to turn about an axis in its plane perpendicular to the field direction. When the coil is turned slightly and released, it oscillates about its stable equilibrium with a frequency of 2.0 s^{-1} . What is the moment of inertia of the coil about its axis of rotation?

ADDITIONAL EXERCISES

- 6.22** A long straight horizontal cable carries a current of 2.5 A in the direction 10° south of west to 10° north of east. The magnetic meridian of the place happens to be 10° west of the geographic meridian. The earth's magnetic field at the location is 0.33 G, and the angle of dip is zero. Locate the line of neutral points (Ignore the thickness of the cable)?
- 6.23** A telephone cable at a place has four long straight horizontal wires carrying a current of 1.0 A in the same direction east to west. The earth's magnetic field at the place is 0.39 G, and the angle of dip is 35° . The magnetic declination is nearly zero. What are the resultant magnetic fields at points 4.0 cm below the cable?
- 6.24** A compass needle free to turn in a horizontal plane is placed at the centre of circular coil of 30 turns and radius 12 cm. The coil is in a vertical plane making an angle of 45° with the magnetic meridian. When the current in the coil is 0.35 A, the needle points west to east.
- Determine the horizontal component of the earth's magnetic field at the location.
 - The current in the coil is reversed, and the coil is rotated about its vertical axis by an angle of 90° in the anticlockwise sense looking from above. Predict the direction of the needle. Take the magnetic declination at the place to be zero.
- 6.25** A magnetic dipole is under the influence of two magnetic fields. The angle between the field directions is 60° , and one of the fields has a magnitude of $1.2 \times 10^{-2} \text{ T}$. If the dipole comes to stable equilibrium at an angle of 15° with this field, what is the magnitude of the other field?
- 6.26** A monoenergetic (18 keV) electron beam initially in the horizontal direction is subject to a horizontal magnetic field of 0.40 G normal to the initial direction. Estimate the up or down deflection of the beam over a distance of 30 cm ($m_e = 9.11 \times 10^{-31} \text{ kg}$, $e = 1.60 \times 10^{-19} \text{ C}$). [Note: Data in this exercise are so chosen that the answer will give you an idea of the effect of earth's magnetic field on the motion of the electron beam from the electron gun, to the screen in a TV set].

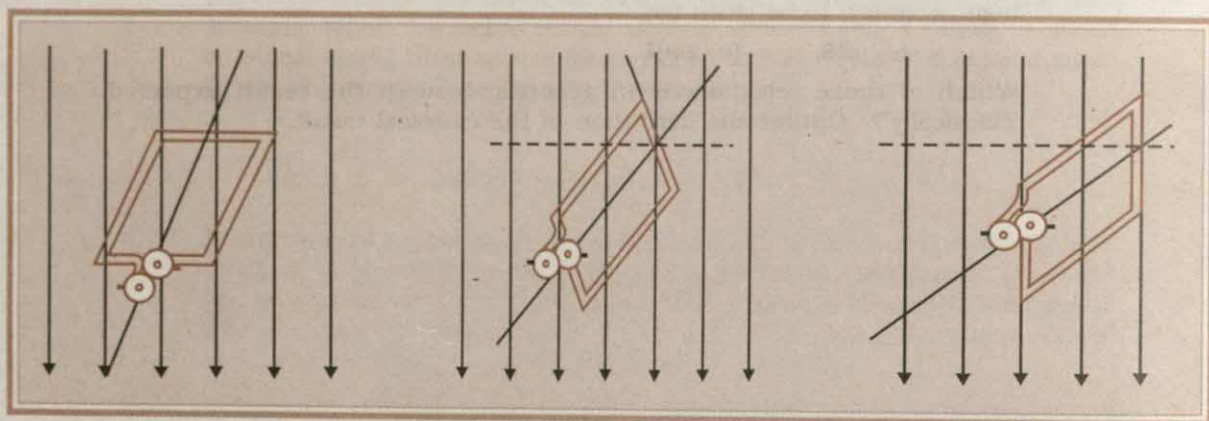
- 6.27** A sample of paramagnetic salt contains 2.0×10^{24} atomic dipoles each of dipole moment $1.5 \times 10^{-23} \text{ J T}^{-1}$. The sample is placed under a homogeneous magnetic field of 0.84 T, and cooled to a temperature of 4.2 K. The degree of magnetic saturation achieved is equal to 15%. What is the total dipole moment of the sample for a magnetic field of 0.98 T and a temperature of 2.8 K? (Assume Curie's law).
- 6.28** A Rowland ring of mean radius 15 cm has 3500 turns of wire wound on a ferromagnetic core of relative permeability 800. What is the magnetic field **B** in the core for a magnetising current of 1.2 A?
- 6.29** The magnetic moment vectors μ_s and μ_l associated with the intrinsic spin angular momentum **S** and orbital angular momentum **l**, respectively, of an electron are predicted by quantum theory (and verified experimentally to a high accuracy) to be given by:

$$\mu_s = -(e/m)\mathbf{S}, \quad \mu_l = -(e/2m)\mathbf{l}$$

Which of these relations is in accordance with the result expected 'classically'? Outline the derivation of the classical result.

CHAPTER SEVEN

ELECTROMAGNETIC INDUCTION



7.1 INTRODUCTION

The word 'induce' means to prevail, or to give rise to. We have seen examples of induction earlier. Under appropriate conditions, a negatively charged rod brought close to a conductor can induce positive charges on the latter. While studying thermoelectricity we have seen that the difference in temperatures can induce an electric current to flow through two dissimilar metals which have been attached to form a closed loop. The present chapter is concerned with a phenomenon of great depth and beauty, namely, the induction of an electromotive force (emf) in a conductor by a changing magnetic field.

We have seen in Chapter 5 that moving charges and currents produce a magnetic field. One may then ask: can moving magnets produce electric currents? Does nature permit such a symmetrical relation between electricity and magnetism? The answer is a resounding yes! The experiments of Michael Faraday in England and of Joseph Henry in USA, conducted around 1830 demonstrated conclusively that changing magnetic fields produced electric fields and electric currents. We will describe some of these experiments in the next section. The phenomenon is appropriately called electromagnetic induction — magnetism inducing electricity.

When Faraday first made public his discovery that relative motion between a bar magnet and a wire loop produced a small current in the latter, he was asked, "What is the use of it?" His reply was: "What is the use of a new born baby?" The phenomenon of electromagnetic induction is not merely of theoretical or academic interest. Imagine a world where there is no electricity — no lights, no trains and no personal computers. The pioneering experiments of Faraday and Henry have led directly to the development of modern day generators and transformers. Civilisation as it exists today owes a great deal to the discovery of electromagnetic induction.



Michael Faraday [1791-1867]

Michael Faraday grew up in London as one of the ten children of a blacksmith. He was apprenticed as a bookbinder but he read widely on electricity and magnetism. He also prepared copious notes on science which he sent to Humphry Davy, a leading authority on heat and thermodynamics. Davy was so impressed with him that he appointed him as his permanent assistant at the Royal Institution.

Faraday made numerous contributions to science, viz., the discovery of electromagnetic induction, the laws of electrolysis, benzene, and the fact that the plane of polarisation is rotated in an electric field. He is also credited with the invention of the electric motor, the electric generator and the transformer. He is widely regarded as the greatest experimental scientist of the nineteenth century.



Joseph Henry [1797-1878]

American experimental physicist professor at Princeton University and first director of the Smithsonian Institution. He made important improvements in electromagnets by winding coils of insulated wire around iron pole pieces, and invented an electromagnetic motor and a new, efficient telegraph. He discovered self-induction and investigated how currents in one circuit induce currents in another.

7.2 THE EXPERIMENTS OF FARADAY AND HENRY

The discovery and understanding of electromagnetic induction was based on a long series of experiments carried out by Faraday and Henry. We shall now describe some of these experiments.

7.2.1 Experiment 1

Figure 7.1 shows a conducting coil connected to a galvanometer. When a bar magnet is pushed towards the coil, the galvanometer deflects. This remarkable fact implies that a current has been set up in the coil in the absence of a battery. The deflection stops when the magnet is made stationary and reverses while the magnet is being pulled away from the coil. If one uses a south pole end of the magnet instead of the north, the experiment works as mentioned above except that the galvanometer deflections are reversed. Further, the experiment works just as well when the coil is moved towards the magnet. What matters is the relative motion of the coil and the magnet. The direction of the current in the loop in Fig. 7.1(a) will be explained later in Section 7.4.

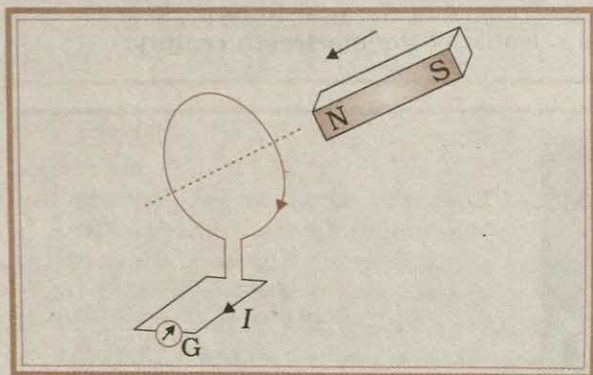


Fig. 7.1(a) When the bar magnet is moved towards the coil, the galvanometer G deflects.

7.2.2 Experiment 2

Figure 7.1(b) shows a coil wound on a cylindrical support and connected to a battery instead of the bar magnet of Fig. 7.1(a). This current carrying coil is moved towards the test coil which is connected to a galvanometer. A similar effect is observed. Once again the galvanometer deflects even in the absence of a source of emf such as a battery. The deflection is greater if the cylindrical support of Fig. 7.1(b) is made of soft iron.

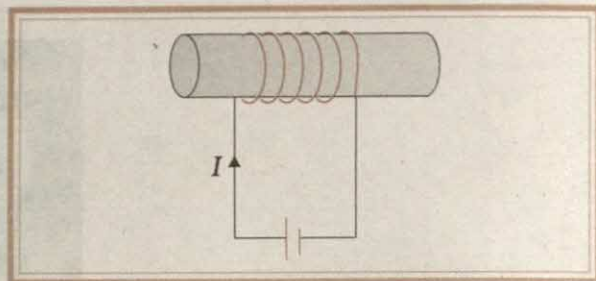


Fig. 7.1(b) A coil wound on a cylindrical support and carrying a current I . This arrangement can replace the bar magnet of Fig. 7.1(a).

7.2.3 Experiment 3

The above two experiments involved relative motion between a magnet and a coil. Faraday showed that this relative motion is not an absolute requirement. Figure 7.2 shows two coils wound on a cylindrical support. The two coils are not connected to each other. When the tapping key* in coil 2 is closed, the galvanometer in coil 1 deflects for a very short time. A plausible explanation for this phenomenon is as follows: it takes some time for the current (and the consequent magnetic field) to establish itself to its maximum value. This time varying magnetic field induces a current in coil 1. We find that when the circuit is opened, there is once again a momentary deflection of the galvanometer but it is in the opposite direction. Further, Faraday found that if the cylindrical support is made of iron the deflections are larger.

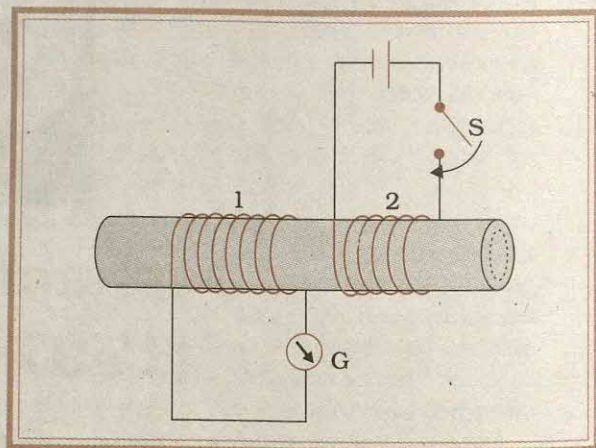


Fig. 7.2 Two coils wound independently on a cylindrical support. When the tapping key S in coil 2 is closed, there is a momentary deflection of the galvanometer G in coil 1.

* A tapping key has an inbuilt spring arrangement so that it can be pressed down to make a momentary connection.

Example 7.1 Consider Experiment 2.

(a) What would you do to obtain a large deflection of the galvanometer? (b) How would you demonstrate the presence of an induced current in the absence of a galvanometer?

Answer

- (a) To obtain a large deflection one would use a rod made of soft iron, a powerful battery so that one gets a large current, and move the arrangement swiftly towards the test coil.
- (b) Replace the galvanometer by a small bulb, the kind one finds in a flash light. The relative motion between the two coils will cause the bulb to glow momentarily and thus demonstrate the presence of an induced current. In experimental physics one must learn to innovate. Michael Faraday who is ranked as one of the best experimentalists ever, was legendary for his innovative skills. ◀

7.3 FARADAY'S LAWS OF INDUCTION**7.3.1 Magnetic Flux**

Faraday's great insight lay in discovering a simple mathematical relation to explain the large series of experiments he carried out on electromagnetic induction. However, before we can state and appreciate his laws, we must get familiar with the notion of magnetic flux. The magnetic flux is defined in the same way as the electric flux in Gauss's law (Chapter 1, Section 1.12). For a uniform magnetic field \mathbf{B} crossing the plane of area \mathbf{A} at an angle with the normal to the plane,

$$\begin{aligned}\Phi_B &= \mathbf{B} \cdot \mathbf{A} \\ &= BA \cos \theta\end{aligned}\quad (7.1)$$

where the notion of the area as a vector has been discussed earlier with the help of Figs. 1.21 and 1.22 in Chapter 1. The Eq. (7.1) can be extended to include non-uniform fields and curved surfaces,

$$\Phi_B = \sum_{\text{all}} \mathbf{B} \cdot d\mathbf{A} \quad (7.2)$$

where 'all' stands for summation over all the elements comprising the surface. For our purposes Eq. (7.1) will suffice. The SI unit of flux is Weber (Wb). Occasionally, it is denoted by T m^2 (tesla meter-square). The magnetic flux is a scalar.

7.3.2 Faraday's Laws

Faraday deduced that a common factor in the experiments described in the previous section is the change in magnetic flux. His two laws are stated as follows:

- (i) *Whenever there is a change of magnetic flux through a circuit, there will be an induced emf and this will last as long as the change persists.*
- (ii) *The magnitude of the induced emf is equal to the time rate of change of magnetic flux.*

Mathematically, the induced emf ε is given by,

$$\varepsilon = - \frac{d\Phi_B}{dt} \quad (7.3)$$

The negative sign indicates the direction of ε and hence the current in a closed loop. We shall discuss it in the next section. If the circuit consists of N turns with associated fluxes $\Phi_1, \Phi_2, \dots, \Phi_N$ then

$$\varepsilon = - \sum_{i=1}^N \frac{d\Phi_i}{dt}$$

A case commonly encountered is of a tightly wound coil or a coil of N turns wound on a rod with the same cross sectional area. In this case, the flux Φ is the same through each winding. Thus,

$$\varepsilon = -N \frac{d\Phi_B}{dt} = - \frac{d(N\Phi_B)}{dt} \quad (7.4)$$

where the second expression is more general. The flux Φ_B can be altered in a variety of ways by: moving a magnet (Fig. 7.1), by a time varying magnetic field (Fig. 7.2), by changing the shape of the coil, that is, by shrinking it or stretching it, etc.

Example 7.2 A square coil of side 10 cm is placed in the east-west plane. A magnetic field of 0.1 T is set up in 0.7 s and in the north-east direction through the coil. The coil has a resistance of 0.7Ω . What is the magnitude of the induced emf and current?

Answer The angle made by the area vector of the coil with the magnetic field is $\theta = 45^\circ$. From Eq. (7.1) the maximum flux is

$$\begin{aligned}\Phi_{\text{max}} &= BA \cos \theta \\ &= \frac{0.1 \times 10^{-2}}{\sqrt{2}} \text{ Wb}\end{aligned}$$

This flux is set up in 0.7 s. From Faraday's law [Eq. (7.3)]; the induced emf has a magnitude,

$$|\varepsilon| = \frac{\Delta\Phi_B}{\Delta t} = \frac{(\Phi_{\text{max}} - 0)}{\Delta t}$$

$$= \frac{10^{-3}}{\sqrt{2} \times 0.7} = 1 \text{ mV}$$

And the magnitude of the current is

$$i = \frac{\varepsilon}{R} = \frac{10^{-3}}{0.7} = 1.4 \text{ mA}$$

Note that: (i) The emf which is generated is quite small. The emphasis on the number of turns N is made to increase the emf. For a tightly wound 100 turn coil $\varepsilon = 0.1 \text{ V}$. However, on account of the corresponding increase in the resistance the current may not show a dramatic increase. (ii) The earth's magnetic field too produces a flux through the coil. But it is a steady field which does not change within the time span of the experiment.

7.4 LENZ'S LAW

In 1824 the German physicist Heinrich Lenz (1804 - 1865) deduced a law, now known as Lenz's law which describes the polarity of the induced emf in a clear concise fashion. The statement of Lenz's law is:

The polarity of the induced emf is such that it tends to produce a current which opposes the change that produces it.

We can understand this law by examining Experiment 1 of Section 7.2.1. In Fig. 7.1(a), we see that the north pole is being pushed towards the coil. The induced current produces a magnetic field. We have seen in Chapter 5 that a current-carrying coil has an associated magnetic dipole moment. The polarity of this dipole is given by the right-hand rule. What Lenz's law implies is that the north pole of this induced current-carrying coil faces and hence repels the north pole of the bar magnet which is being pushed towards it. Another way to interpret this law is as follows: the bar magnet's motion increases the flux through the coil. The induced current opposes and reduces this magnetic flux.

The above discussion can be further clarified with the help of Fig. 7.3. When the north pole of a magnet approaches a coil, the induced current in the coil flows in a direction such that the coil presents its 'north pole' to the bar magnet. This is depicted in Fig. 7.3(a). The magnetic field of the induced current tends to repel the approaching bar magnet. When the north pole of the bar magnet moves away from the coil, the coil presents its 'south pole' to the bar magnet [Fig. 7.3(b)]. The coil tends to attract the receding bar magnet.

A little reflection should convince us of the correctness of Lenz's law. Suppose that the induced current was in the opposite direction to the one depicted in Fig 7.1(a) or Fig. 7.3(a). Then the coil presents its south face to the incoming north pole of the magnet. The bar magnet will be sucked in through the coil at an ever increasing acceleration. A gentle push towards the magnet will initiate a process that will continuously increase its velocity and kinetic energy. By a suitable arrangement one could construct a perpetual motion machine! This violates the law of conservation of energy. What actually happens is that the energy required to set up the induced current is dissipated into Joule heating. We will quantify this notion in Section 7.6.

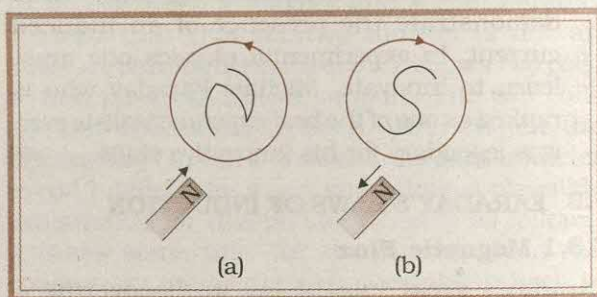


Fig. 7.3 Illustration of Lenz's law.

Lenz's law is usually understood in terms of closed circuits and *induced currents*. If the circuit were open, we should think in terms of what would happen if it were closed and thus determine the polarity of the induced emf.

7.5 MOTIONAL EMF AND FARADAY'S LAW

It is possible to understand Faraday's law when the conductor is moving and the magnetic field is uniform and time independent. The conductor has free charges and the Lorentz force acting on them can cause a current to be set up. We now quantify this notion with a concrete example. Figure 7.4 shows a rectangular conductor in which one arm PQ is free to move. It is placed in a uniform magnetic field which is perpendicular to the plane of the conductor. The arm PQ is moved inwards with a speed v . The flux through the loop is Blx , so the induced emf is,

$$\varepsilon = -\frac{d}{dt}(Blx) = -Bl \frac{dx}{dt} = Blv \quad (7.5)$$

where $dx/dt = -v$ is used. The induced emf Blv is termed as **motional emf**. It is possible to understand the motional emf expression in

Eq. (7.5) by invoking the Lorentz force acting on the free charge carriers of arm PQ. Consider a charge q at P and at the instant when the motion of the arm PQ is initiated with the speed v . The Lorentz force on this charge is qvB in magnitude, and its direction is towards Q. The work done in moving a charge towards Q is,

$$W = qvBl$$

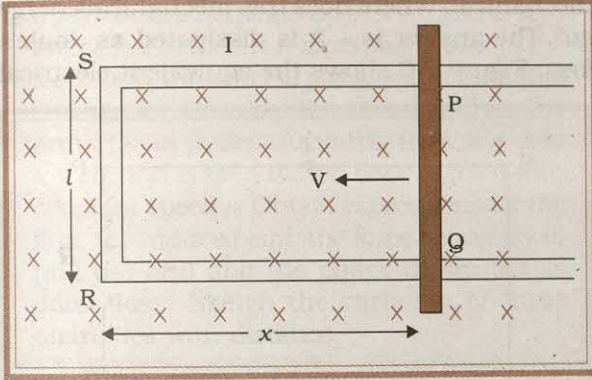


Fig. 7.4 The arm PQ is moved inwards thus changing the area of the rectangular loop. This induces a current I as shown.

The emf is the work done per unit charge,

$$e = W/q$$

$$= Blv$$

which is identical to Eq. (7.5). We stress that our presentation is not wholly rigorous. But it does help us to understand the basis of Faraday's law when the conductor is moving in a uniform and time independent magnetic field.

On the other hand, it is not obvious how an emf is induced when the conductor is stationary and the magnetic field is changing – a fact which Faraday verified by numerous experiments. The force on a charge is given in complete generality by,

$$\mathbf{F} = q(\mathbf{E} + \mathbf{v} \times \mathbf{B}) \quad (7.6)$$

Since $\mathbf{v} = 0$, any force on the charge must arise from the electric field term \mathbf{E} . We are led to a conclusion that a time varying magnetic field will generate an electric field. In Chapter 5, we learnt that charges in motion (currents) could exert a force/torque on a stationary bar magnet. Conversely, a bar magnet in motion (or more generally, a changing magnetic field) can exert a force on the stationary charge. This is the fundamental significance of the Faraday's discovery. Electricity and magnetism are related.

Example 7.3 A conducting rod 1m length moves with a frequency of 50 rev/s, with one end at the center and the other end at the circumference of a circular metallic ring of radius 1m, about an axis passing through the center of the coil perpendicular to the plane of the coil. A constant magnetic field parallel to the axis is present everywhere. What is the emf between the center and the metallic ring? Given that $B = 1$ T.

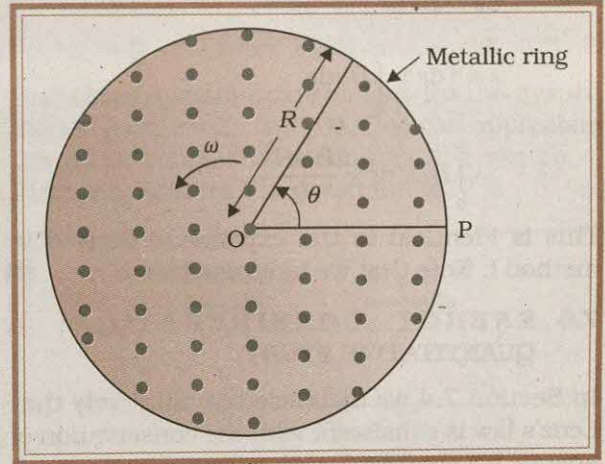


Fig. 7.5 Figure for Example 7.3.

Answer

Method 1 There are no steady currents in this example. Separation of charges takes place. To calculate the emf, we can imagine a closed loop by connecting the center with any point on the circumference, say P, with a resistor. The potential difference across the resistor is then equal to the induced emf and equals $B \times$ (rate of change of area of loop). If θ is the angle between the rod and the radius of the circle at P at time t , the area of the arc formed by the rod and the radius at P is:

$$\pi R^2 \times \frac{\theta(t)}{2\pi} = \frac{1}{2} R^2 \theta(t)$$

where R is the radius of the circle. Hence the induced emf is

$$B \times \frac{d}{dt} \left[\frac{1}{2} R^2 \theta(t) \right]$$

$$= \frac{1}{2} BR^2 \frac{d\theta(t)}{dt}$$

$$= \frac{1}{2} \times 1.0 \times (1^2) \times 50 \times 2\pi$$

$$= 157 \text{ V}$$

[Note : $\frac{d\theta}{dt} = \omega = 2\pi\nu$]

Method II We use the notion of motional emf. The magnitude of the emf generated along a length of the rod as it moves at right angles to the magnetic field is, from Eq (7.5),

$$d\varepsilon = Bvdr$$

$$\varepsilon = \int d\varepsilon = \int_0^R Bvdr$$

$$= \int_0^R B\omega r dr = \frac{B\omega R^2}{2}$$

This is identical to the expression derived in method I. Note that we have used $v = \omega r$.

7.6 ENERGY CONSIDERATION: A QUANTITATIVE STUDY

In Section 7.4 we discussed qualitatively that Lenz's law is consistent with the conservation of energy. We shall explore this aspect further with a concrete example.

Let R be the resistance of movable arm PQ of the rectangular conductor in Fig. 7.4. We assume that the remaining arms QR , RS and SP have negligible resistance compared to R . Thus, the overall resistance of the rectangular loop is R and this does not change as PQ is dragged inwards. The current I in the loop is,

$$I = \frac{\varepsilon}{R}$$

$$= \frac{Blv}{R} \quad (7.7)$$

On account of the presence of the magnetic field, there will be a force on the arm PQ . This force - $I(\mathbf{L} \times \mathbf{B})$, is directed outwards in the direction opposite to the velocity of the rod in consonance with Lenz's law. The magnitude of this force is,

$$F = I l B$$

$$= \frac{B^2 l^2 v}{R}$$

where we have used Eq. (7.7).

Alternatively, the arm PQ is being pushed with a constant speed v . The power required to do this is,

$$P = Fv$$

$$= \frac{B^2 l^2 v^2}{R} \quad (7.8)$$

The agent that does this work is probably mechanical. Where does this mechanical energy go? The answer is - it is dissipated as Joule's loss. Figure 7.6 shows the equivalent electrical

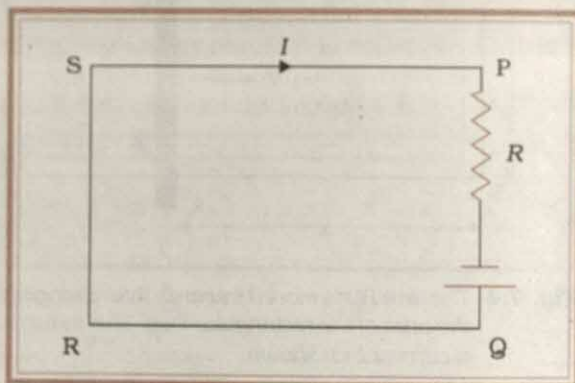


Fig. 7.6 The electrical circuit equivalent of the electromagnetic set up of Fig. 7.4.

circuit corresponding to the set up of Fig. 7.4. The Joule loss is

$$P_J = I^2 R$$

$$= \left(\frac{Blv}{R} \right)^2 R$$

$$= \frac{B^2 l^2 v^2}{R}$$

which is identical to Eq. (7.8).

Thus, mechanical energy which was needed to move the arm PQ is converted to electrical energy (the induced emf) and then to thermal energy.

There is an interesting relationship between the charge flow through the circuit and the change in the magnetic flux. From Faraday's law, we have that the magnitude of the induced emf is,

$$|\varepsilon| = \frac{\Delta\Phi_B}{\Delta t}$$

However,

$$|\varepsilon| = IR = \frac{\Delta Q}{\Delta t} R$$

Thus,

$$\Delta Q = \frac{\Delta \Phi_B}{R}$$

Example 7.4 This example is the reverse of the example discussed in Sections 7.5 and 7.6. The arm PQ of the rectangular conductor is moved from $x = 0$, outwards. The uniform magnetic field extends up to $x = b$ and is zero for $x > b$. Further, it is a perpendicular to the loop. Only the arm PQ possesses substantial resistance r . Consider the situation when the arm PQ is pulled outwards from $x = 0$ to $x = 2b$, and is then moved back to $x = 0$ with constant speed v . Obtain expressions for the flux, the induced emf, the force necessary to pull the arm and the power dissipated as Joule loss. Sketch the variation of these quantities with distance.

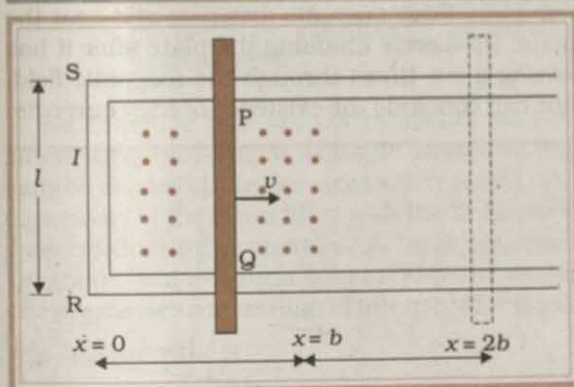


Fig. 7.7

Answer Let us first consider the forward motion from $x = 0$ to $x = 2b$

The flux Φ_B is

$$\begin{aligned}\Phi_B &= Blx & 0 \leq x < b \\ &= Blb & b \leq x < 2b\end{aligned}$$

The induced emf is,

$$\begin{aligned}\varepsilon &= -\frac{d\Phi_B}{dt} \\ &= -Blv & 0 \leq x < b \\ &= -0 & b \leq x < 2b\end{aligned}$$

When the induced emf is non-zero, the current I is (in magnitude)

$$I = \frac{Blv}{r}$$

The force required to keep the arm PQ in constant motion is Ilb . Its direction is to the left. In magnitude

$$\begin{aligned}F &= \frac{B^2 l^2 v}{r} & 0 \leq x < b \\ &= 0 & b \leq x < 2b\end{aligned}$$

The Joule heating loss is

$$\begin{aligned}P_J &= I^2 R \\ &= \frac{B^2 l^2 v^2}{r} & 0 \leq x < b \\ &= 0 & b \leq x < 2b\end{aligned}$$

One obtains similar expressions for the inward motion from $x = 2b$ to $x = 0$. One can appreciate the whole process by examining the sketch of various quantities displayed in Fig. 7.8.

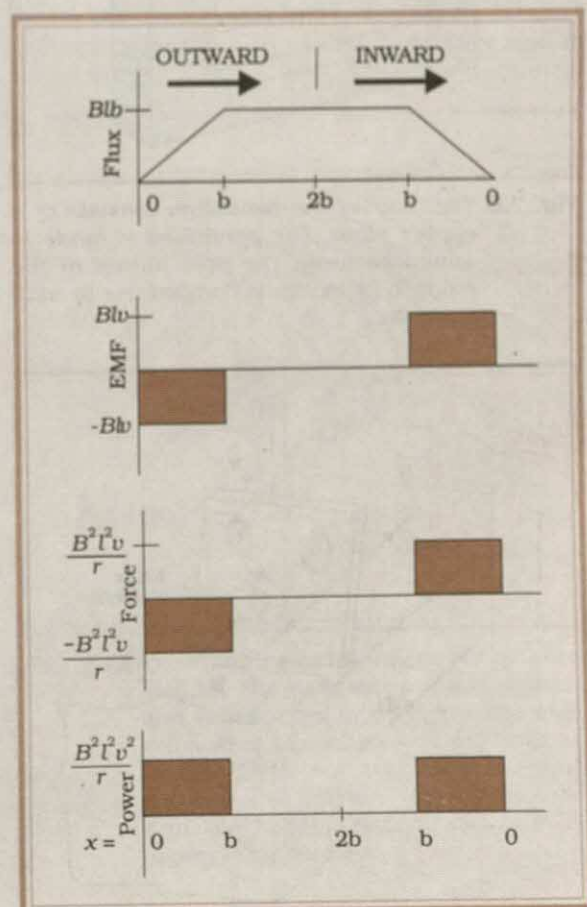


Fig. 7.8 Sketch of various quantities connected with Example 7.4. Note that the magnetic force on the arm PQ is inward while it is being pulled outwards and vice-versa.

7.7 EDDY CURRENTS

Consider the apparatus shown in Fig. 7.9. A copper plate is made to swing between the pole pieces of a magnet like a simple pendulum. It is found that the motion is damped and in a little while the plate comes to a halt in the magnetic field.

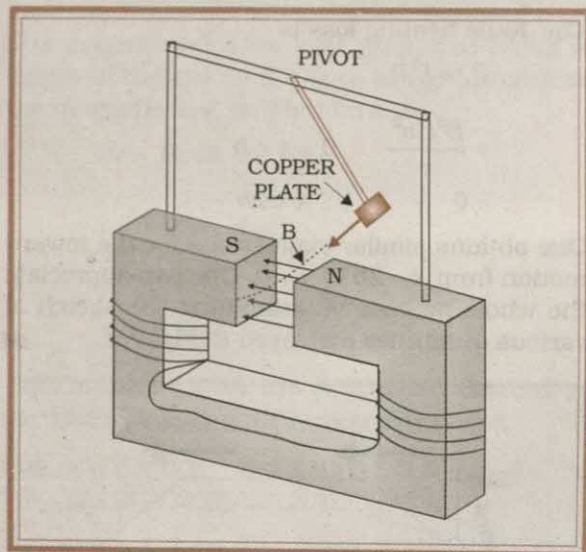


Fig. 7.9 The 'bob' of the pendulum consists of a copper plate. The pendulum is made to swing between the pole pieces of the magnet. Its motion is damped due to eddy currents.

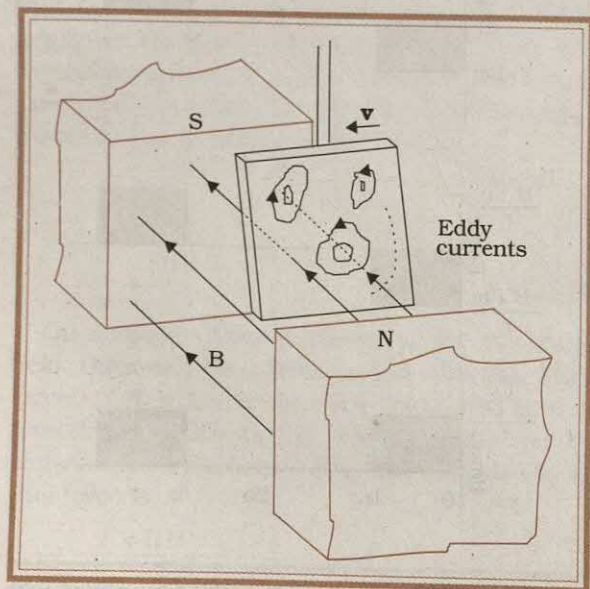


Fig. 7.10 A schematic view of the eddy currents which break the motion of the pendulum shown in Fig. 7.9.

We can explain the above mentioned phenomenon on the basis of electromagnetic induction. Fig. 7.10 shows that because of flux change a current is induced in the plate. These currents which seek a path of the least resistance form irregularly shaped loops. Their directions however are not random, but dictated by Lenz's law. In Fig. 7.10, they are anticlockwise as the plate swings into the field and clockwise as the plate swings out of the field. Such currents are called **eddy currents**.

The presence of eddy currents can also be inferred by making rectangular slots in the copper plate. This arrangement is shown in Fig. 7.11. The slot interrupts the flow of eddy current. Only eddy currents enclosing a very small area are then possible. From the expression for the retarding force F just before Eq. (7.8), we note that $F \propto l^2$, i.e., F scales with the area. Thus, punching holes or slots reduce electromagnetic damping. Eddy currents dissipate and heat the plate. By merely touching the plate after it has swung a few times through the magnetic field, one can conclude the existence of eddy currents.

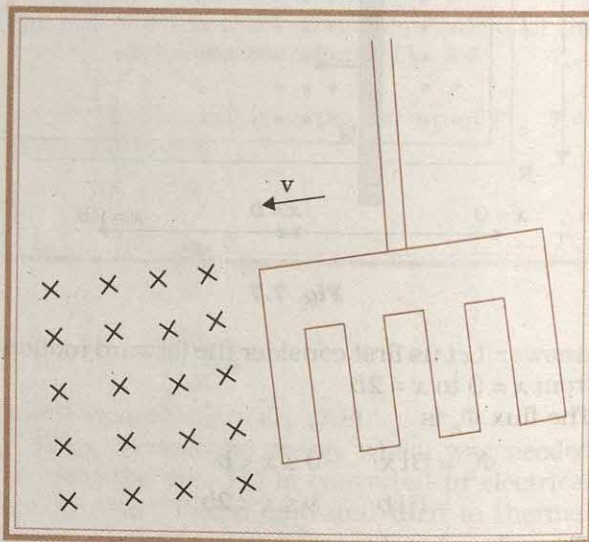


Fig. 7.11 Cutting slots in the copper plate reduces the effect of eddy currents.

Eddy currents are undesirable since they dissipate energy in the form of heat. To reduce the loss, the conducting parts are built of a large number of thin layers separated by an insulating material like lacquer. This arrangement, known as lamination, increases the resistance of the possible paths to the flow of the eddy currents.

Eddy currents may also be used to advantage. (i) The braking systems in many modern trains make use of eddy currents. Note that the force due to the eddy current is proportional to speed ($F \propto v$). Thus, the braking effect is smooth. (ii) As Fig. 7.12 indicates, eddy currents may be used for electromagnetic shielding. (iii) They are also used in speedometers and (iv) in induction furnaces.

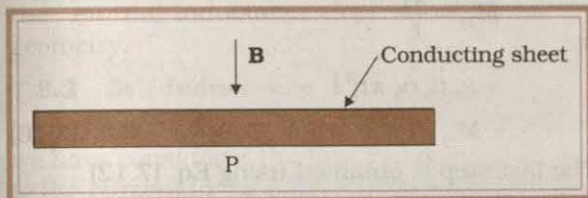


Fig. 7.12 A magnetic field B is suddenly switched on and directed downwards. Eddy currents are generated in the sheet and this change in the magnetic field is only partially detected at P . The higher the conductivity of the sheet, the better the shielding of the transient magnetic field.

We conclude this section by mentioning an interesting biomedical example which should only be carried out under strict safety conditions. The cavity of the eye is filled with the fluid which has conducting properties. A large transient magnetic field (~ 1 tesla field at 60 hertz) across the eye causes a sensation of intense brightness.

7.8 INDUCTANCE

Consider two coils. The current through one coil sets up a magnetic field both in the second coil as well through itself. Inductance is basically a measure of the ratio of the flux to the current I . We have seen that the magnetic field $B \propto I$ and thus the flux.

$$\Phi_B \propto I$$

If the geometry of the coils is not varying with time then,

$$\frac{d\Phi_B}{dt} \propto \frac{dI}{dt}$$

The constant of proportionality is called inductance. We shall see that it depends on the geometry of the coil and intrinsic material properties. This aspect is akin to capacitance which for a parallel plate capacitor depends on the plate area and plate separation (geometry) and the dielectric constant K of the interposing

medium (intrinsic material property); or like resistance which depends on the geometry, say, length and cross-sectional area of the wire and resistivity which is an intrinsic material property.

Inductance is a scalar. It has the dimensions of flux divided by current as seen from the expressions of the previous paragraph. In terms of the fundamental quantities its dimensions are: $[ML^2T^{-2}A^{-2}]$. The SI unit of inductance is called henry and is denoted by H . It is named in honor of Joseph Henry who discovered electromagnetic induction in USA independently of Faraday in England.

7.8.1 Mutual Inductance

Consider Fig. 7.13 which shows two long co-axial solenoids each of length l . We denote the radius of the inner solenoid S_1 by r_1 and the number of turns per unit length by n_1 . The corresponding quantities for the outer solenoid S_2 are r_2 and n_2 respectively.

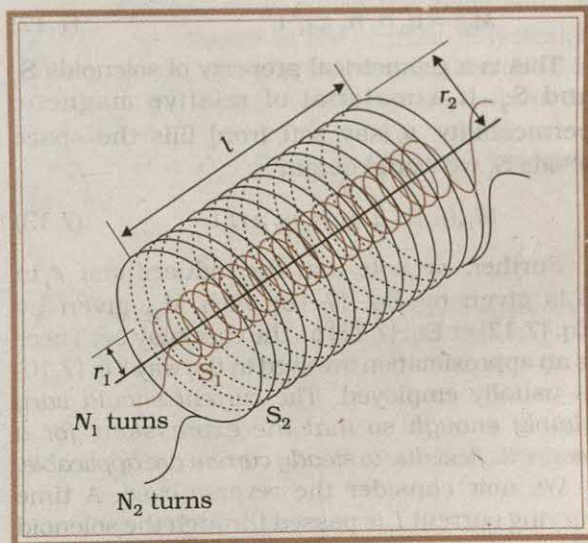


Fig. 7.13 Two long co-axial solenoids of same length l . The inner solenoid has radius r_1 and n_1 turns per unit length. The outer solenoid S_2 has radius r_2 and n_2 turns per unit length ($n_1 l = N_1$; $n_2 l = N_2$). The figure will serve to define mutual inductance and self inductance and obtain expressions for them.

We set up a time varying current I_2 through S_2 . This sets up a time varying magnetic flux through S_1 which we designate by Φ_1 . The mutual inductance is the constant of

proportionality given by,

$$\Phi_1 = M_{12} I_2 \quad (7.9)$$

M_{12} is called the mutual inductance of the circuit 1 with respect to the circuit 2. It is sometimes also referred to as the coefficient of mutual induction.

From Faraday's law the induced emf in S_1 is,

$$\begin{aligned} \varepsilon_1 &= -\frac{d\Phi_1}{dt} \\ &= -M_{12} \frac{dI_2}{dt} \end{aligned} \quad (7.10)$$

For this simple co-axial solenoids it is possible to calculate M_{12} . The magnetic field due to the current I_2 is $\mu_0 n_2 I_2$. The flux Φ_1 through S_1 is,

$$\begin{aligned} \Phi_1 &= (\pi r_1^2)(\mu_0 n_2 I_2)(n_1 l) \\ &= \mu_0 n_1 n_2 \pi r_1^2 l I_2 \end{aligned} \quad (7.11)$$

where note that $n_1 l$ is the total number of turns of solenoid S_1 . Thus, from Eq. (7.9) and Eq. (7.11),

$$M_{12} = \mu_0 n_1 n_2 \pi r_1^2 l \quad (7.12)$$

This is a geometrical property of solenoids S_1 and S_2 . If a material of relative magnetic permeability μ_r (say soft iron) fills the space inside S_1 we would obtain

$$M_{12}(\mu_r) = \mu_r \mu_0 n_1 n_2 \pi r_1^2 l \quad (7.13)$$

Further, we note that the induced emf ε_1 in S_1 is given by Eq. (7.10) with M_{12} given by Eq. (7.12) or Eq. (7.13) as the case may be. There is an approximation involved in the way Eq. (7.10) is usually employed. *The current should vary slowly enough so that the expressions for a magnetic field due to steady current are applicable.*

We now consider the reverse case. A time varying current I_1 is passed through the solenoid S_1 and the associated flux through S_2 is Φ_2 ,

$$\Phi_2 = M_{21} I_1 \quad (7.14)$$

M_{21} is called the mutual inductance of the circuit 2 with respect to the circuit 1.

From Faraday's law, the induced emf ε_2 in S_2 is,

$$\begin{aligned} \varepsilon_2 &= -\frac{d\Phi_2}{dt} \\ &= -M_{21} \frac{dI_1}{dt} \end{aligned} \quad (7.15)$$

Once again for this simple case of co-axial solenoids, it is possible to calculate M_{21} . The flux

due to S_1 is confined solely inside S_1 since the solenoids are very long. There is no magnetic field outside S_1 , due to the current in S_1 . Thus, Φ_2 is

$$\Phi_2 = (\pi r_1^2)(\mu_0 n_1 I_1)(n_2 l)$$

where $n_2 l$ is the total number of turns of S_2 . From Eq. (7.14),

$$\begin{aligned} M_{21} &= \frac{\Phi_2}{I_1} \\ &= \mu_0 n_1 n_2 \pi r_1^2 l \\ &= M_{12} \end{aligned} \quad (7.16)$$

The last step is obtained using Eq. (7.12).

It is not always easy to calculate the mutual inductance. However, it is helpful to keep three aspects in mind: (1) Mutual inductance depends on the geometry of the two circuits. (2) Mutual inductance depends on the intrinsic magnetic property of the material, for example, the relative permeability. (3) The equality,

$$M_{12} = M_{21} \quad (7.17)$$

holds. We have demonstrated this equality for long co-axial solenoids. However, the relation is far more general even though we do not prove it in the present text. The equality is very useful as we see in the next example.

Example 7.5 Two circular coils, one of small radius r_1 and the other of very large radius r_2 are placed co-axially with centers coinciding. Obtain the mutual inductance of the arrangement.

Answer Let a current I_2 flow through the outer circular coil. The field at the center of the coil is $B_2 = \mu_0 I_2 / 2r_2$. Since the second co-axially placed coil has very small radius, B_2 may be considered constant over its cross-sectional area. Hence,

$$\begin{aligned} \Phi_1 &= \pi r_1^2 B_2 \\ &= \frac{\mu_0 \pi r_1^2}{2r_2} I_2 \\ &= M_{12} I_2 \end{aligned}$$

Hence,

$$M_{12} = \frac{\mu_0 \pi r_1^2}{2r_2}$$

From Eq. (7.17)

$$M_{12} = M_{21} = \frac{\mu_0 \pi r_1^2}{2r_2}$$

Note that it would have been difficult to calculate the flux through the bigger coil of the non-uniform field due to the current in the smaller coil and hence the mutual inductance M_{21} . The equality $M_{12} = M_{21}$ is helpful. Note also that mutual inductance depends solely on the geometry.

7.8.2 Self-Inductance

In the discussion in the previous sub-section, we have considered the flux in one solenoid due to the current in the other. Consider the general case of currents flowing simultaneously in two nearby coils. The flux linked with one coil will be due to the sum of two fluxes which exist independently. The law of superposition applies to magnetic fields. For example, Eq. (7.9) would generalise to,

$$\Phi_1 = M_{11} I_1 + M_{12} I_2 \quad (7.18)$$

Therefore, using Faraday's law,

$$\varepsilon_1 = -M_{11} \frac{dI_1}{dt} - M_{12} \frac{dI_2}{dt}$$

M_{11} is the self-inductance and is written as L_1 . It is also sometimes called the coefficient of self-induction.

$$\varepsilon_1 = -L_1 \frac{dI_1}{dt} - M_{12} \frac{dI_2}{dt}$$

The self-induced emf will exist even if there is a single coil. This single coil will have a self inductance which we term L . Thus

$$\varepsilon = -L \frac{dI}{dt} \quad (7.19)$$

One can understand self-inductance by examining the isolated circuit of Fig. 7.14.

The circuit of Fig. 7.14 consists of a single-turn coil through which the current is gradually increased by lowering the resistance of the rheostat. The magnetic flux linked with the coil increases with time. The increasing flux induces an emf in the circuit. By Lenz's law, the direction of the induced current is opposite to the direction of the conventional current. The increase in current in the circuit is not instantaneous, but it is gradual. This effect is called self-induction, since the changing flux in the circuit arises from the current in the very same circuit.

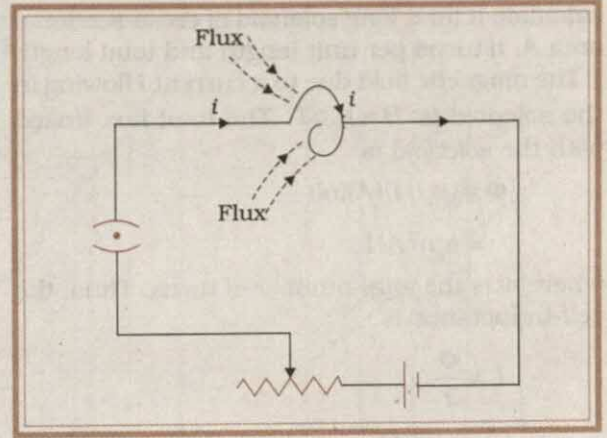


Fig. 7.14 An isolated circuit. When the switch is closed and current is increased by lowering the rheostat resistance, the flux (shown by dashed lines) linked with the single-turn coil increases. This induces an emf in the coil which by Lenz's law is in a direction opposite to the battery emf.

The induced emf is called the 'back emf'. It opposes any change in the current. Physically, the self inductance plays the role of inertia. It is the electromagnetic analogue of mass in mechanics.

The rate of work (power) needed in an electric circuit is,

$$\frac{dW}{dt} = |\varepsilon| I$$

If we ignore the resistive losses and consider only inductive effect, then using Eq. (7.19),

$$\frac{dW}{dt} = LI \frac{dI}{dt}$$

one can write,

$$2I \frac{dI}{dt} = \frac{d(I^2)}{dt}$$

Thus,

$$\frac{dW}{dt} = \frac{1}{2} L \frac{d(I^2)}{dt}$$

or the energy required to build up the current is,

$$W = \frac{1}{2} L I^2 \quad (7.20)$$

The expression reminds us of the kinetic energy expression $mv^2/2$.

It is possible to calculate the self-inductance for circuits with simple geometries. Let us

calculate it for a long solenoid of cross-sectional area A , n turns per unit length and total length l . The magnetic field due to a current I flowing in the solenoid is $B = \mu_0 n I$. The total flux linked with the solenoid is

$$\Phi = (\mu_0 n I)(A)(nl)$$

$$= \mu_0 n^2 A l I$$

where nl is the total number of turns. Thus, the self-inductance is,

$$L = \frac{\Phi}{I} = \mu_0 n^2 A l \quad (7.21)$$

If we fill the inside of the solenoid with a material of high relative magnetic permeability μ_r (example soft iron), then,

$$L(\mu_r) = \mu_r \mu_0 n^2 A l \quad (7.22)$$

The self-inductance of the coil depends on its geometry and on μ_r .

Example 7.6 (a) What is the magnetic energy stored in the solenoid? (b) How does this magnetic energy compare with the electrostatic energy stored in a capacitor?

Answer

(a) From Eq. (7.20), the magnetic energy is

$$\begin{aligned} U_m &= \frac{1}{2} L I^2 \\ &= \frac{1}{2} L \left(\frac{B}{\mu_0 n} \right)^2 \quad (\text{since } B = \mu_0 n I) \\ &= \frac{1}{2} (\mu_0 n^2 A l) \left(\frac{B}{\mu_0 n} \right)^2 \quad [\text{from Eq. (7.21)}] \\ &= \frac{1}{2 \mu_0} B^2 A l \end{aligned}$$

(b) The magnetic energy per unit volume V is,

$$\begin{aligned} u_m &= \frac{U_m}{V} \\ &= \frac{U_m}{A l} \\ &= \frac{B^2}{2 \mu_0} \end{aligned} \quad (7.23)$$


If we make the substitution,

$$\frac{1}{\mu_0} \rightarrow \epsilon_0$$

$\mathbf{B} \rightarrow \mathbf{E}$ (Electric field),

we obtain the electrostatic energy stored per unit volume in a parallel plate capacitor.

$$u_e = \frac{1}{2} \epsilon_0 E^2 \quad (7.24)$$

In both the cases energy is proportional to the square of the field strength. Eqs. (7.23) and (7.24) have been derived for special cases: a solenoid and a parallel plate capacitor, respectively. But they are general and valid for any region of space in which a magnetic field or electric field exists. 

7.9 AC GENERATOR

The phenomenon of electromagnetic induction has been technologically exploited in many ways. An exceptionally important application is the generation of alternating currents (AC). The modern AC generator with a typical output capacity of 100 MW (million watts) is a highly evolved machine. In this section, we shall describe the basic principles behind this machine. The Yugoslav inventor Nicola Tesla is credited with the development of the machine. The principle underlying the AC generator may be simply stated as follows: *the AC generator is a device which converts mechanical energy to electrical energy on the basis of electromagnetic induction.*

In its simplest form, the AC generator consists of an N turn loop of wire of cross-sectional area A rotating in a uniform magnetic field \mathbf{B} . In Fig. 7.15 we show the arrangement at different instants of time. At any given instant the loop makes an angle θ with the field. This angle is

$\theta = 0, \theta, \frac{\pi}{2}$ in Fig. 7.15(a), (b), (c), respectively.

Let us choose to count the rotation from the instant when $\theta = 0$. Then if the angular velocity ω of the loop is constant,

$$\theta = \omega t$$

The flux at instant t is,

$$\begin{aligned} \Phi_m(t) &= N \mathbf{B} \cdot \mathbf{A} \\ &= N B A \cos \theta \\ &= N B A \cos(\omega t) \end{aligned}$$

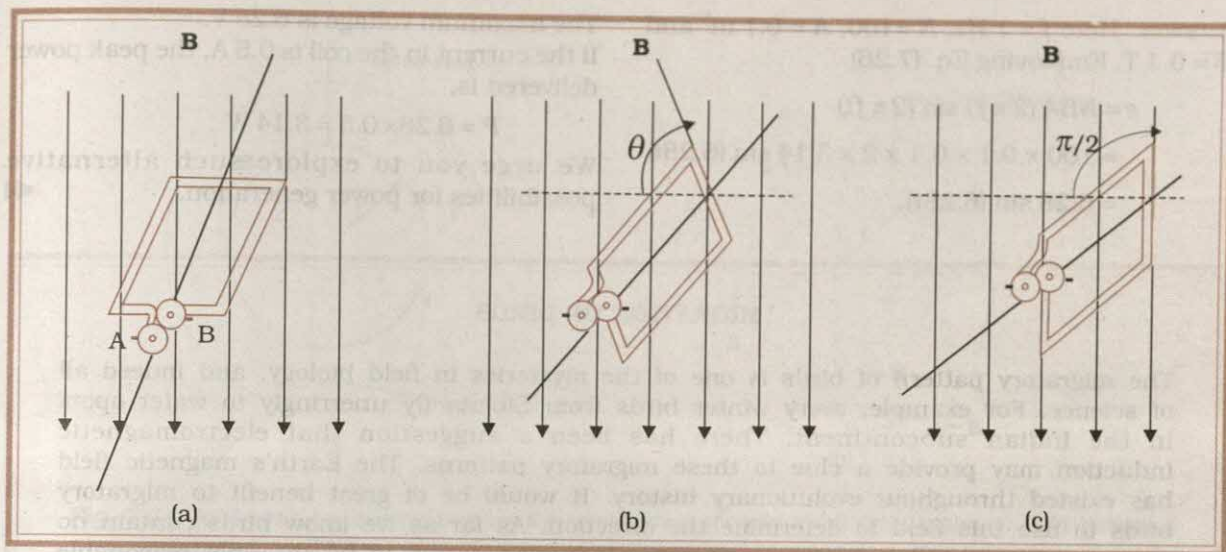


Fig. 7.15 The principle of an AC generator is illustrated. A coil is rotated in a uniform magnetic field, thus inducing motional emf. The coil is shown such that its area vector makes different orientations with \mathbf{B} : (a) $\theta = 0$; (b) $\theta = \theta$ (general); (c) $\theta = \pi/2$.

From Faraday's law of induction

$$\begin{aligned}\varepsilon &= -\frac{d\Phi}{dt} \\ &= NBA\omega \sin(\omega t) \\ &= V_m \sin(\omega t)\end{aligned}$$

where the maximum generated emf is,

$$V_m = NBA\omega$$

This sinusoidal variation of the emf is displayed in Fig. 7.16.

If the coil makes f revolutions per second then,

$$\omega = 2\pi f$$

and $V_m = 2\pi f NBA$

The next question is how do you tap into this emf? This query has spawned a technology of its own. In its simplest form, metallic strings are connected to the two ends of the loop which are co-axial with the axis of rotation of the loop. These rings are called slip rings and are shown in Fig. 7.15 (a) as A and B. A and B are connected to two brushes which are stationary carbon pieces. The brushes are connected to an external circuit.

In commercial generators, the mechanical energy required for rotation is provided by water falling from a height. Hence, dams are constructed. These are called hydro-electric generators. Alternatively, water is heated to produce steam using coal or sources. The steam at high pressure produces the rotation. These

are called thermal generators. If instead of coal a nuclear fuel is used we get nuclear power (Chapter 14). Modern day generators produce power as high as 500 MW, i.e., one can light up 5 million 100 W bulbs! In most generators the coils are held stationary and it is electromagnets which are rotated. The frequency of rotation is 50 Hz in India. In certain countries such as USA, it is 60Hz.

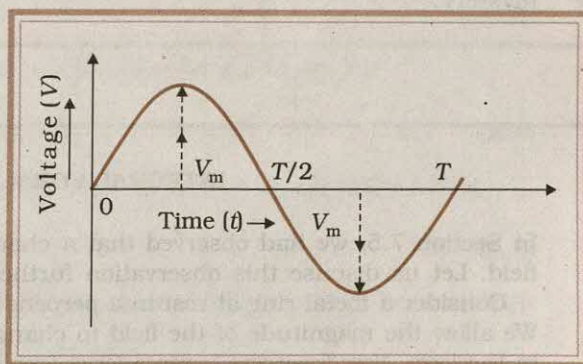


Fig. 7.16 The motional emf produced by the rotating coil of Fig. 7.14. The time period $T = 2\pi/\omega$.

Example 7.7 Siddharth peddles a stationary bicycle at one revolution per second. The pedals are attached to 100 turn coil of area 0.1 m^2 and placed in a uniform magnetic field of 0.1 Tesla . What is the maximum voltage generated in the coil?


Answer Here $f = 1 \text{ Hz}$; $N = 100$, $A = 0.1 \text{ m}^2$ and $B = 0.1 \text{ T}$. Employing Eq. (7.26)

$$\begin{aligned}\epsilon &= NBA (2\pi f) \sin(2\pi ft) \\ &= 100 \times 0.1 \times 0.1 \times 2 \times 3.14 \sin(6.28t) \\ &= 6.28 \sin(6.28t).\end{aligned}$$

The maximum voltage is 6.28 V .

If the current in the coil is 0.5 A , the peak power delivered is,

$$P = 6.28 \times 0.5 = 3.14 \text{ W}$$

We urge you to explore such alternative possibilities for power generation. 

MIGRATION OF BIRDS

The migratory pattern of birds is one of the mysteries in field biology, and indeed all of science. For example, every winter birds from Siberia fly unerringly to water spots in the Indian subcontinent. There has been a suggestion that electromagnetic induction may provide a clue to these migratory patterns. The Earth's magnetic field has existed throughout evolutionary history. It would be of great benefit to migratory birds to use this field to determine the direction. As far as we know birds contain no ferromagnetic material. So electromagnetic induction seems to be the only reasonable mechanism to determine direction. Consider the optimal case where the magnetic field \mathbf{B} , the velocity of the bird \mathbf{v} , and two relevant points of its anatomy are separated by a distance l , and all three are mutually perpendicular. From the formula for motional emf Eq. (7.5),

$$\epsilon = Blv$$

Taking $B = 4 \times 10^{-5} \text{ T}$, $l = 2 \text{ cm}$ wide, and $v = 10 \text{ m/s}$, we obtain,

$$\begin{aligned}\epsilon &= 4 \times 10^{-5} \times 2 \times 10^{-2} \times 10 = 8 \times 10^{-6} \text{ V} \\ &= 8 \mu\text{V}\end{aligned}$$

This extremely small potential difference suggests that our hypothesis is of doubtful validity. Certain kinds of fish are able to detect small potential differences. However, in these fish, special cells have been identified which detect small voltage differences. In birds no such cells have been identified. Thus, the migration patterns of birds continues to remain a mystery.

INTEGRAL FORM OF FARADAY'S LAW

In Section 7.5, we had observed that a changing magnetic field will give rise to an electric field. Let us discuss this observation further.

Consider a metal ring at rest in a perpendicular magnetic field \mathbf{B} as shown in Fig. 7.24(a). We allow the magnitude of the field to change at the rate dB/dt . This will induce a current in the loop. Now from the Lorentz force formula we know that a magnetic field will not exert a force on a stationary charge. To explain the existence of the induced current, we must assume that the **changing magnetic field induces an electric field \mathbf{E}** . This field will exert a force $q\mathbf{E}$ on a mobile charge q of the ring. As the charge moves a distance $d\mathbf{l}$ along the ring, the work done on it by the induced field is,

$$\Delta W = q \mathbf{E} \cdot d\mathbf{l}$$

The electric field induced at various points along the metal ring is tangential and has the same magnitude by symmetry. Thus, the work done in moving the charge once around the loop is,

$$W = qE (2\pi r)$$

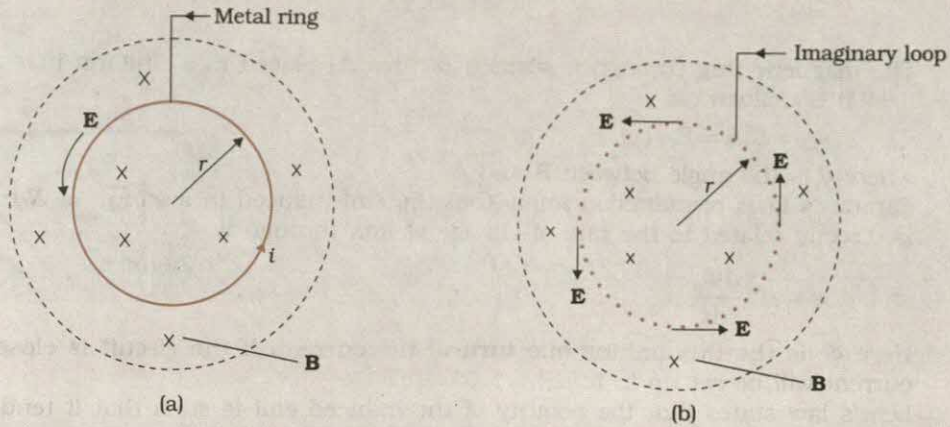


Fig. 7.24 (a) An induced current i appears in a metal ring of radius r when the magnetic field increases with time. (b) The metal ring is removed. An induced (tangential) electric field exists along the imaginary loop of radius r even in the absence of the ring.

where r is the radius of the ring. The emf is defined as the work done per unit charge. Thus

$$\varepsilon = E (2 \pi r) \quad (\text{A.1})$$

For a more general loop we may write Eq. (A.1) as

$$\varepsilon = \oint \mathbf{E} \cdot d\mathbf{l} \quad (\text{A.2})$$

where the integral is taken over the length of the loop. From Faraday's law of induction,

$$\varepsilon = -\pi r^2 \frac{dB}{dt} \quad (\text{A.3})$$

or more generally,

$$\varepsilon = -\frac{d\Phi_B}{dt} \quad (\text{A.4})$$

Eqs. (A.2) and (A.4) permit us to write down an integral form of Faraday's law,

$$\oint \mathbf{E} \cdot d\mathbf{l} = -\frac{d\Phi_B}{dt} \quad (\text{A.5})$$

For the simple case of the metal ring shown in Fig. 7.24(a), Eqs. (A.1) and (A.3) imply,

$$E = \frac{r}{2} \frac{dB}{dt} \quad (\text{A.6})$$

Eqs. (A.5) and (A.6) tell us how exactly the changing magnetic field is related to an electric field.

It is important to realise that an induced electric field appears even in the absence of metal ring. This is shown in Fig. 7.24(b). The **betatron**, which is a machine used to accelerate electrons, works on this principle.

The induced electric field is **not** related to static electric charges. It forms closed loops (like magnetic field lines) and does not originate and terminate at positive and negative charges, respectively. You **cannot** associate a potential with it. The induced electric field is **not conservative**. In contrast, the electrostatic field is conservative and its integral over a closed loop [right hand side of Eq. (A.5)] is strictly zero.

SUMMARY

1. The magnetic flux through a surface (of area **A**) placed in a uniform magnetic field **B** is defined as,

$$\Phi_B = \mathbf{B} \cdot \mathbf{A} = BA \cos \theta$$

where θ is the angle between **B** and **A**.

2. Faraday's laws of induction imply that the emf induced in a circuit of N turns is directly related to the rate of change of flux through it,

$$\varepsilon = -N \frac{d\Phi_B}{dt}$$

Here Φ_B is the flux linking one turn of the current. If the circuit is closed a current will be set up in it.

3. Lenz's law states that the polarity of the induced emf is such that it tends to produce a current which opposes the change that produces it. The negative sign in the expression for Faraday's law indicates this fact.
4. When a conducting rod of length l is moved in a magnetic field of magnitude B with a velocity v such that the arrangement is mutually perpendicular, then the induced emf (called motional emf) is

$$\varepsilon = Blv$$

5. Changing magnetic fields can set up current loops in nearby conductors. Such loops are irregularly shaped. They dissipate energy as heat. Such current loops are called eddy currents.
6. Inductance is a measure of the induced flux due to current I . It is the ratio Φ/I .
7. A changing current in a coil (coil 2) can induce an emf in a nearby coil (coil 1). This relation is given by,

$$\varepsilon_1 = -M_{12} \frac{dI_2}{dt}$$

The quantity M_{12} is called mutual inductance of coil 1 due to coil 2. One can similarly define M_{21} . There exists a general equality,

$$M_{12} = M_{21}$$

which we prove only for a special case in this text.

8. When a current changes in a coil, it induces a back emf in the same coil. This self induced emf is given by,

$$\varepsilon = -L \frac{dI}{dt}$$

L is called the inductance of the coil. It is a measure of the opposition of a coil to a change of current through it.

9. For a long solenoid, whose core consists of a magnetic material of permeability μ_r ,

$$L = \mu_r \mu_0 n^2 Al$$

Here A is the cross-section area of the solenoid, l its length and n the number of turns per unit length.

10. In an AC generator, mechanical energy is converted to electrical energy on the basis of electromagnetic induction. If an N turn coil of area A is rotated at f revolutions per second in a uniform magnetic field B , then the motional emf produced is sinusoidal:

$$\varepsilon = NBA (2\pi f) \sin (2\pi ft)$$

where we have assumed that at time $t = 0$ s, the coil is perpendicular to the field.

Quantity	Symbol	Units	Dimensions	Remarks
Magnetic Flux	Φ_B	Wb (weber)	$[ML^2T^{-2}A^{-1}]$	Magnetic Field Area
EMF	ε	V (volt)	$[ML^2T^{-3}A^{-1}]$	$\varepsilon = -d\Phi_B/dt$
Mutual Inductance	M	H (henry)	$[ML^2T^{-2}A^{-2}]$	$\varepsilon_1 = -M_{12}dI_2/dt$
Self Inductance	L	H (henry)	$[ML^2T^{-2}A^{-2}]$	$\varepsilon = -LdI/dt$

POINTS TO PONDER

1. Electricity and magnetism are intimately related. In the early part of the nineteenth century the experiments of Oersted, Ampere and others established that moving charges (currents) produce a magnetic field. Somewhat later, around 1830 the experiments of Faraday and Henry demonstrated that a moving magnet can induce an electric current. One might then ask: Are all the fundamental forces that we know, the gravitational, the electromagnetic, the weak and strong nuclear forces, inter-related?
2. An emf is induced in a circuit where the magnetic flux is changing even if the circuit is open. Closing the circuit results in a current flow through the circuit.
3. Lenz's law is consistent with the principle of energy conservation. If the current were to flow in a direction opposite to that dictated by Lenz's law, then perpetual motion machines would be possible!
4. The induced emf in Faraday's law does not have a chemical or electrostatic origin. The work done in taking a charge around the closed loop is not zero. If \mathbf{E} is the electric field related to the induced emf,

$$\oint \mathbf{E} \cdot d\mathbf{l} \neq 0$$

where the integral is taken over the closed loop of the circuit. The induced electric field \mathbf{E} is not a conservative field. The closed loop need not be an actual material circuit. We can imagine any closed curve in space. The integral of induced electric field over the curve is non-zero.

5. The motional emf discussed in Section 7.5 can be argued independently from Faraday's law using the Lorentz force on moving charges. However, even if the charges are stationary [and the $q(\mathbf{v} \times \mathbf{B})$ term of the Lorentz force is not operative], an emf is nevertheless induced in the presence of a time varying magnetic field. Thus, moving charges in static field and static charges in a time-varying field seem to be symmetric situation for Faraday's law. This gives a tantalising hint on the relevance of the principle of relativity for Faraday's law.
6. Inductance may be viewed as electrical inertia. The analogue of self inductance in mechanics is mass.
7. Both self-inductance and mutual inductance depend on (i) geometry of the circuit and (ii) intrinsic material property such as the magnetic permeability. In this sense they bear a similarity to capacitance and resistance.
8. The capacitance, resistance, inductance, and the diode (to be described later in this book) constitute the four passive elements of an electrical circuit. Indeed they are the four alphabets of electrical/electronic engineering and technology.

EXERCISES

7.1 Figure 7.17 below shows planar loops of different shapes moving out of or into a region of a magnetic field which is directed normal to the plane of the loop away from the reader. Determine the direction of induced current in each loop using Lenz's law.

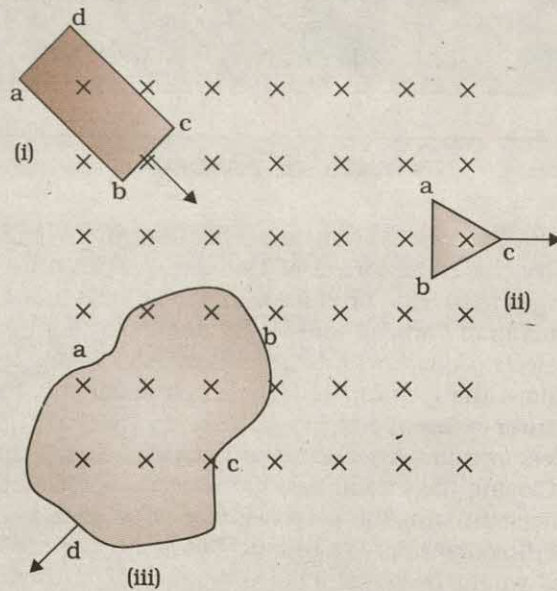
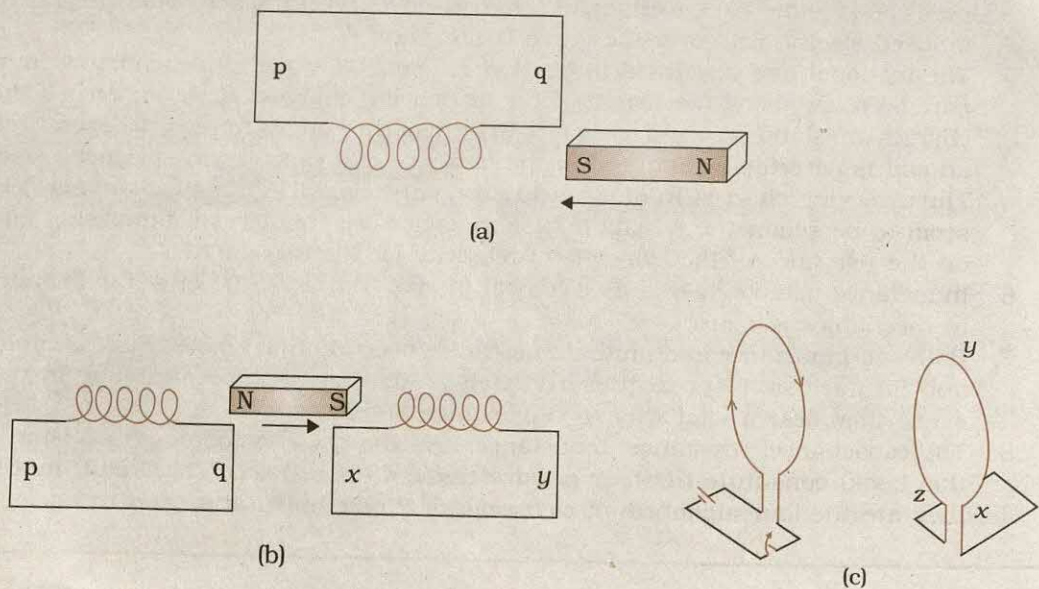


Fig. 7.17

7.2 Predict the direction of induced current in the situations described by the following Fig. 7.18(a) to (f).



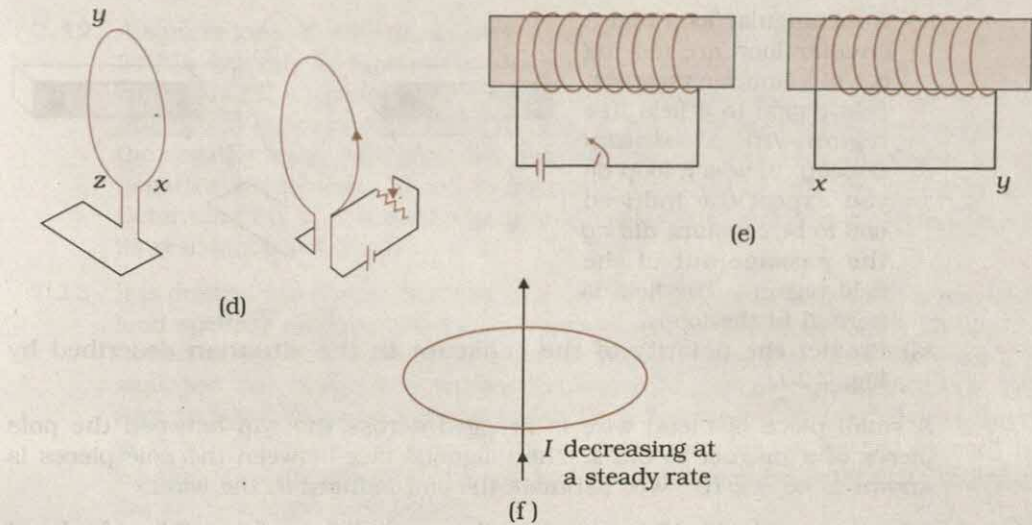


Fig. 7.18

7.3 Use Lenz's law to determine the direction of induced current in the situations described by Fig. 7.19:

- (a) A wire of irregular shape turning into a circular shape;
 (b) A circular loop being deformed into a narrow straight wire.

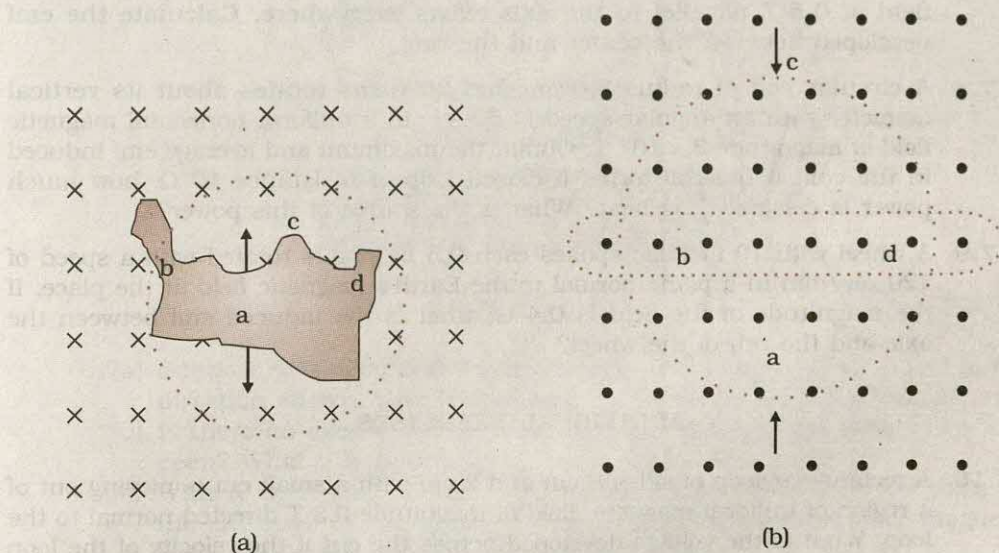


Fig. 7.19

7.4 Answer the following questions:

- (a) A conducting loop is held stationary normal to the field between the NS poles of a fixed permanent magnet. By choosing a magnet sufficiently strong, can we hope to generate current in the loop?
- (b) A closed conducting loop moves normal to the electric field between the plates of a large capacitor. Is a current induced in the loop when it is (i) wholly inside the capacitor (ii) partially outside the plates of capacitor? The electric field is normal to the plane of the loop.

- (c) A rectangular loop and a circular loop are moving out of a uniform magnetic field region to a field free region with a *constant velocity*. In which loop do you expect the induced emf to be constant *during* the passage out of the field region? The field is normal to the loops.

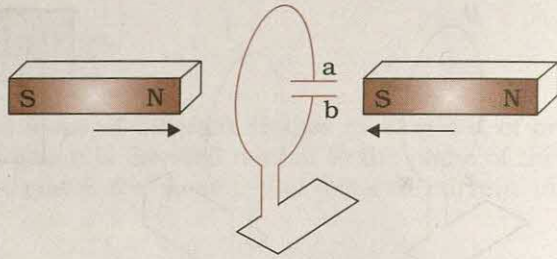


Fig. 7.20

- (d) Predict the polarity of the capacitor in the situation described by Fig. 7.20.

- 7.5** A small piece of metal wire is dragged across the gap between the pole pieces of a magnet in 0.5 s. The magnetic flux between the pole pieces is known to be 8×10^{-4} Wb. Estimate the emf induced in the wire.
- 7.6** A long solenoid with 15 turns per cm has a small loop of area 2.0 cm^2 placed inside normal to the axis of the solenoid. If the current carried by the solenoid changes steadily from 2 A to 4 A in 0.1 s, what is the induced voltage in the loop while the current is changing?
- 7.7** A 1 m long conducting rod rotates with an angular frequency of 400 s^{-1} about an axis normal to the rod passing through its one end. The other end of the rod is in contact with a circular metallic ring. A constant magnetic field of 0.5 T parallel to the axis exists everywhere. Calculate the emf developed between the center and the ring.
- 7.8** A circular coil of radius 8.0 cm and 20 turns rotates about its vertical diameter with an angular speed of 50 s^{-1} in a uniform horizontal magnetic field of magnitude 3×10^{-2} T. Obtain the maximum and average emf induced in the coil. If the coil forms a closed loop of resistance 10Ω , how much power is dissipated as heat? What is the source of this power?
- 7.9** A wheel with 10 metallic spokes each 0.5 m long is rotated with a speed of 120 rev/min in a plane normal to the Earth's magnetic field at the place. If the magnitude of the field is 0.4 G, what is the induced emf between the axle and the rim of the wheel?

ADDITIONAL EXERCISES

- 7.10** A rectangular loop of sides 8 cm and 2 cm with a small cut is moving out of a region of uniform magnetic field of magnitude 0.3 T directed normal to the loop. What is the voltage developed across the cut if the velocity of the loop is 1 cm s^{-1} in a direction normal to the (i) longer side, (ii) shorter side of the loop? For how long does the induced voltage last in each case?
(Note: This and some other exercises ignore one important point for simplicity: A magnetic field cannot abruptly change in space from a finite value to zero).
- 7.11** Suppose the loop in Exercise 7.10 is stationary but the current feeding the electromagnet that produces the magnetic field is gradually reduced so that the field decreases from its initial value of 0.3 T at the rate of 0.02 T s^{-1} . If the cut is joined and the loop has a resistance of 1.6Ω , how much power is dissipated by the loop as heat? What is the source of this power?

7.12 A square loop of side 12 cm with its sides parallel to X and Y axes is moved with a velocity of 8 cm s^{-1} in the positive x-direction in an environment containing a magnetic field in the positive z-direction. The field is neither uniform in space nor constant in time. It has a gradient of $10^{-3} \text{ T cm}^{-1}$ along the negative x-direction (that is it increases by $10^{-3} \text{ T cm}^{-1}$ as one moves in the negative x-direction), and it is decreasing in time at the rate of 10^{-3} T s^{-1} . Determine the direction and magnitude of the induced current in the loop if its resistance is $4.5 \text{ m}\Omega$.

7.13 It is desired to measure the magnitude of field between the poles of a powerful loud speaker magnet. A small flat search coil of area 2 cm^2 with 25 closely wound turns, is positioned normal to the field direction, and then quickly snatched out of the field region. Equivalently, one can give it a quick 90° turn to bring its plane parallel to the field direction). The total charge flown in the coil (measured by a ballistic galvanometer connected to coil) is 7.5 mC . The resistance of the coil and the galvanometer is 0.50Ω . Estimate the field strength of magnet.

7.14 Figure 7.21 shows a metal rod PQ resting on the rails AB and positioned between the poles of a permanent magnet. The rails, the rod, and the magnetic field are in three mutual perpendicular directions. A galvanometer G connects the rails through a switch K. Length of the rod = 15 cm , $B = 0.50 \text{ T}$, resistance of the closed loop containing the rod = $9.0 \text{ m}\Omega$.

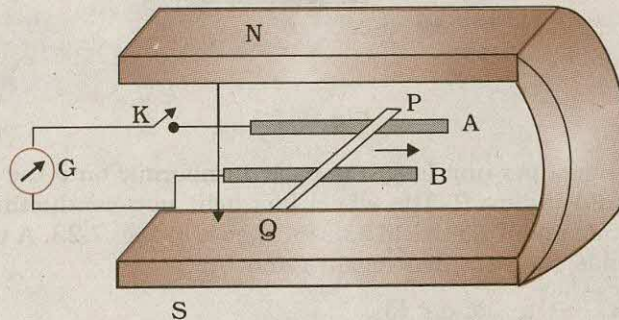


Fig. 7.21

- Suppose K is open and the rod moves with a speed of 12 cm s^{-1} in the direction shown. Give the polarity and magnitude of the induced emf.
- Is there an excess charge built up at the ends of the rods when K is open? What if K is closed?
- With K open and the rod moving uniformly, there is *no net force* on the electrons in the rod PQ even though they do experience magnetic force due to the motion of the rod. Explain.
- What is the retarding force on the rod when K is closed?
- How much power is required (by an external agent) to keep the rod moving at the same speed ($=12 \text{ cm s}^{-1}$) when K is closed? How much power is required when K is open?
- How much power is dissipated as heat in the closed circuit? What is the source of this power?
- What is the induced emf in the moving rod when the permanent magnet is rotated to a vertical position so that the field is parallel to the rails?

7.15 An air-cored solenoid with length 30 cm , area of cross-section 25 cm^2 and number of turns 500 , carries a current of 2.5 A . The current is suddenly

switched off in a brief time of 10^{-3} s. How much is the average back emf induced across the ends of the open switch in the circuit? Ignore the variation in magnetic field near the ends of the solenoid.

- 7.16** A jet plane is travelling west at the speed of 1800 km/h. What is the voltage difference developed between the ends of the wing 25 m long, if the Earth's magnetic field at the location has a magnitude of 5×10^{-4} and the dip angle is 30° .
- 7.17** (a) Obtain an expression for the mutual inductance between a long straight wire and a square loop of side a as shown in Fig. 7.22.
 (b) Evaluate the induced emf in the loop if the wire carries a current of 50 A and the loop has an instantaneous velocity $v = 10 \text{ m s}^{-1}$ at the location $x = 0.2 \text{ m}$, as shown. Take $a = 0.1 \text{ m}$ and assume that the loop has a large resistance.

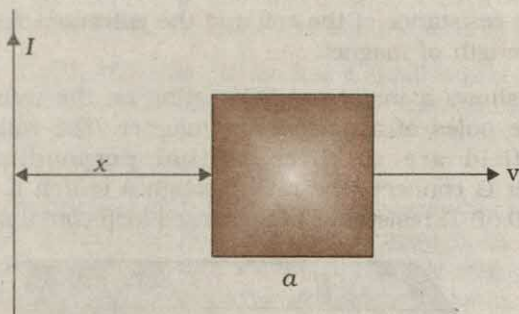


Fig. 7.22

- 7.18** A line charge λ per unit length is pasted uniformly onto the rim of a wheel of mass M and radius R . The wheel has light non-conducting spokes and is free to rotate about a vertical axis as shown in Fig. 7.23. A uniform magnetic field extends over a radial region given by,

$$\mathbf{B} = -B_0 \mathbf{k} \quad (r \leq a; a < R)$$

$$= 0 \quad (\text{otherwise})$$

What is the angular velocity of the wheel when this field is suddenly switched off?

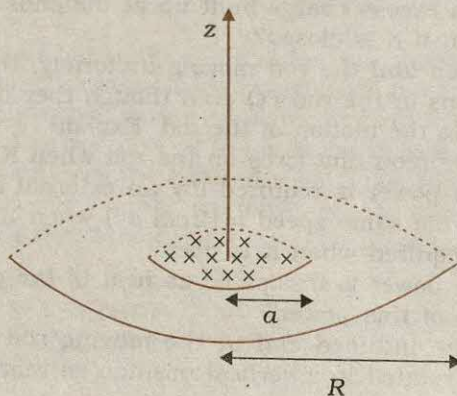
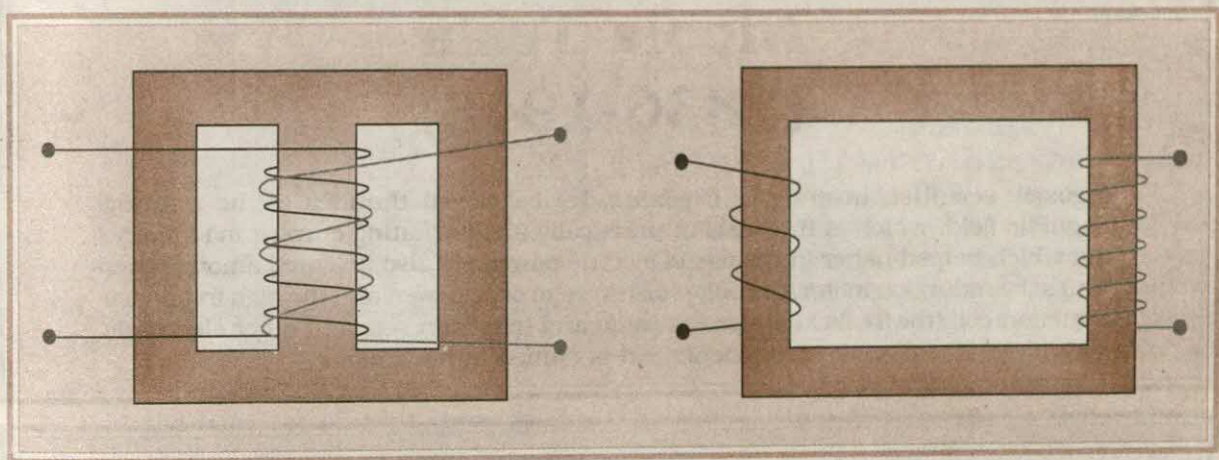


Fig. 7.23

CHAPTER EIGHT

ALTERNATING CURRENT



8.1 INTRODUCTION

We have so far considered direct current (dc) sources and circuits with dc sources. These currents do not change direction with time. But voltages and currents that vary with time are very common. The electric mains supply in our homes and offices is a voltage that varies like a sine function with time. Such a voltage is called alternating voltage (ac voltage) and the current driven by it in a circuit is called the alternating current (ac current)*. Today, most of the electrical devices we use require ac voltage. This is mainly because most of the electrical energy sold by power companies is transmitted and distributed as alternating current. The main reason for preferring use of ac voltage over dc voltage is that ac voltages can be easily and efficiently converted from one voltage to the other by means of transformers. Further, electrical energy in form of ac can also be transmitted economically over long distances. Alternating circuits exhibit characteristics which are exploited in many devices of daily use. For example, whenever we tune our radio to a favourite station, we are taking advantage of a special property of ac circuits – one of many that you will study in this Chapter.

* The phrases 'ac voltage' and 'ac current' are contradictory and redundant, respectively, since they mean, literally, 'alternating current voltage' and 'alternating current current'. Still, the abbreviation ac to designate an electrical quantity displaying simple harmonic time dependence has become so universally accepted that we follow others in its use. Further, 'voltage' – another phrase commonly used means potential difference between two points.



Nicola Tesla (1836-1943)

Yugoslav scientist, inventor and genius. He conceived the idea of the rotating magnetic field, which is the basis of practically all alternating current machinery, and which helped usher in the age of electric power. He also invented among other things the induction motor, the polyphase system of ac power, and the high frequency induction coil (the Tesla coil) used in radio and television sets and other electronic equipment. The SI unit of magnetic field is named in his honour.



George Westinghouse (1846-1914)

A leading proponent of the use of alternating current over direct current. Thus, he came into conflict with Thomas Alva Edison, an advocate of direct current. Westinghouse was convinced that the technology of alternating current was the key to the electrical future. He founded the famous Company named after him and enlisted the services of Nicola Tesla and other inventors in the development of alternating current motors and apparatus for the transmission of high tension current, pioneering in large scale lighting.

8.2 AC VOLTAGE APPLIED TO A RESISTOR

Figure 8.1 shows a resistor connected to a source ε of ac voltage. The symbol for an ac source in a circuit diagram is \ominus . For simplicity, we consider a source which produces sinusoidally varying potential difference across its terminals. Let this potential difference, also called ac voltage, be given by

$$V = V_m \sin \omega t \quad (8.1)$$

where V_m is the amplitude of the oscillating potential difference and ω is its angular

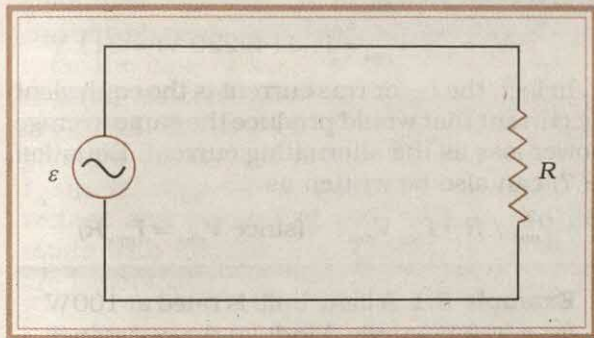


Fig. 8.1 AC voltage applied to a resistor.

frequency.

To find the value of current through the resistor, we apply Kirchhoff's loop rule $\sum \varepsilon(t) = 0$, to the circuit shown in Fig. 8.1 to get

$$V_m \sin \omega t = IR$$

$$\text{or} \quad I = \frac{V_m}{R} \sin \omega t$$

Since R is a constant, we can write this equation as

$$I = I_m \sin \omega t \quad (8.2)$$

where the current amplitude I_m is given by

$$I_m = \frac{V_m}{R} \quad (8.3)$$

Equation (8.3) is just Ohm's law which for resistors works equally well for both ac and dc voltages. The voltage across a pure resistor and the current through it, given by Eqns.(8.1) and (8.2) are plotted as a function of time in Fig. 8.2. Note, in particular that both V and I reach zero, minimum and maximum values at the same time. Clearly, the voltage and current are in phase with each other.

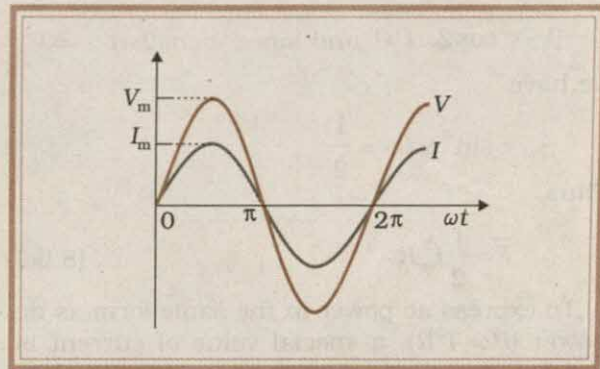


Fig. 8.2 In a pure resistor, the voltage and current are in phase. The minima with zero and maxima occur at the same respective times.

We see that, like the applied voltage, the current varies sinusoidally and has corresponding positive and negative values during each cycle. Thus, the sum of the instantaneous current values over one complete cycle is zero, and the average current is zero. The fact that the average current is zero, however, does not mean that the average power is zero and that there is no dissipation of electrical energy. As you know, joule heating is given by $I^2 R$ and depends on I^2 (which is always positive whether I is positive or negative) and not on I . Thus, there is Joule heating and dissipation of electrical energy when an ac current passes through a resistor.

The instantaneous power dissipated in the resistor is

$$P = I^2 R = I_m^2 R \sin^2 \omega t \quad (8.4)$$

The average value of P over a cycle is*

$$\bar{P} = \langle I^2 R \rangle = \langle I_m^2 R \sin^2 \omega t \rangle \quad (8.5a)$$

where the bar over a letter (here, P) denotes its average value and $\langle \dots \rangle$ denotes taking average of the quantity inside the bracket. Since, I_m^2 and R are constants,

$$\bar{P} = I_m^2 R \langle \sin^2 \omega t \rangle \quad (8.5b)$$

Using the trigonometric identity, $\sin^2 \omega t = \frac{1}{2}(1 - \cos 2\omega t)$, we have $\langle \sin^2 \omega t \rangle$

* Mathematically, the average value of a function $F(t)$ over a period T is given by $\langle F(t) \rangle = \frac{1}{T} \int_0^T F(t) dt$.

$= \frac{1}{2}(1 - \langle \cos 2\omega t \rangle)$ and since $\langle \cos 2\omega t \rangle = 0^*$, we have,

$$\langle \sin^2 \omega t \rangle = \frac{1}{2}$$

Thus,

$$\bar{P} = \frac{1}{2} I_m^2 R \quad (8.5c)$$

To express ac power in the same form as dc power ($P = I^2 R$), a special value of current is used. It is called, *root mean square (rms) or effective current* (Fig. 8.3) and is denoted by I_{rms} . It is defined by

$$I_{rms} = \sqrt{\bar{I}^2} = \sqrt{\frac{1}{2} I_m^2} = \frac{I_m}{\sqrt{2}} = 0.707 I_m \quad (8.6)$$

In terms of I_{rms} , the average power is

$$\bar{P} = \frac{1}{2} I_m^2 R = I_{rms}^2 R \quad (8.7)$$

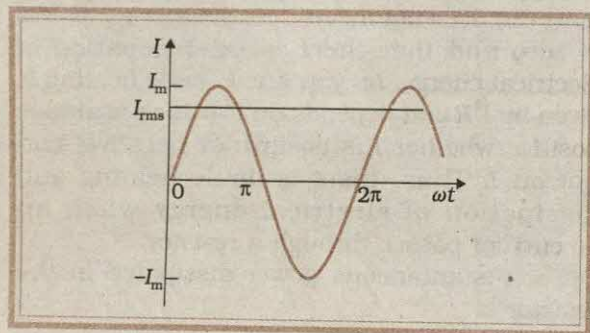


Fig. 8.3 The rms current I_{rms} is related to the peak

current I_m by $I_{rms} = \frac{I_m}{\sqrt{2}} = 0.707 I_m$.

Similarly, we define the *rms voltage or effective voltage* by

$$V_{rms} = \frac{V_m}{\sqrt{2}} = 0.707 V_m \quad (8.8)$$

From Eq. (8.3), we have

$$V_m = I_m R$$

$$\text{or } \frac{V_m}{\sqrt{2}} = \frac{I_m}{\sqrt{2}} R$$

$$\text{or } V_{rms} = I_{rms} R \quad (8.9)$$

Equation (8.9) gives the relation between ac current and voltage and is similar to that in the dc case. This shows the advantage of introducing the concept of rms values. In terms of rms values, the equation for power and relation between current and voltage in ac circuits are essentially the same as those for the dc case.

It is customary to measure and specify rms values for ac quantities. For example, the household line voltage of 220 V is an rms value with a peak voltage of

$$V_m = \sqrt{2} V_{rms} = (1.414)(220 \text{ V}) = 311 \text{ V}$$

In fact, the I_{rms} or rms current is the equivalent dc current that would produce the same average power loss as the alternating current. Equation (8.7) can also be written as

$$\bar{P} = V_{rms}^2 / R = I_{rms} V_{rms} \quad (\text{since } V_{rms} = I_{rms} R)$$

Example 8.1 A light bulb is rated at 100 W for a 220 V supply. Find: (a) the resistance of the bulb; (b) the peak voltage of the source; and (c) the rms current through the bulb.

Answer

(a) We are given $P = 100 \text{ W}$ and $V = 220 \text{ V}$. The resistance of the bulb is

$$R = \frac{V_{rms}^2}{P} = \frac{(220 \text{ V})^2}{100 \text{ W}} = 484 \Omega$$

(b) The peak voltage of the source is

$$V_m = \sqrt{2} V_{rms} = 311 \text{ V}$$

(c) Since, $P = I_{rms} V_{rms}$

$$I_{rms} = \frac{P}{V_{rms}} = \frac{100 \text{ W}}{220 \text{ V}} = 0.45 \text{ A}$$

8.3 REPRESENTATION OF AC CURRENT AND VOLTAGE BY ROTATING VECTORS — PHASORS

In the previous section, we saw that the current through a resistor is in phase with the ac voltage. But this is not so in the case of an

$$\langle \cos 2\omega t \rangle = \frac{1}{T} \int_0^T \cos 2\omega t dt = \frac{1}{T} \left[\frac{\sin 2\omega t}{2\omega} \right]_0^T = \frac{1}{2\omega T} [\sin 2\omega T - 0] = 0.$$

inductor, a capacitor or a combination of these circuit elements. In order to show phase relationship between voltage and current in an ac circuit, we use the notion of PHASORS. The analysis of an ac circuit is facilitated by the use of a phasor diagram. A phasor* is a vector which rotates about the origin with angular speed ω , as shown in Fig. 8.4. The vertical components of phasors \mathbf{V} and \mathbf{I} represent the sinusoidally varying quantities V and I . The magnitudes of phasors \mathbf{V} and \mathbf{I} represent the amplitudes or the peak values V_m and I_m of these oscillating quantities. Figure 8.4(a) shows the voltage and current phasors and their relationship at time t_1 for the case of an ac source connected to a resistor i.e., corresponding to the circuit shown in Fig. 8.1. The projection of voltage and current phasors on vertical axis, i.e., $V_m \sin \omega t$ and $I_m \sin \omega t$, respectively represent the value of voltage and current at that instant. As they rotate with frequency ω , curves in Fig. 8.4(b) are generated.

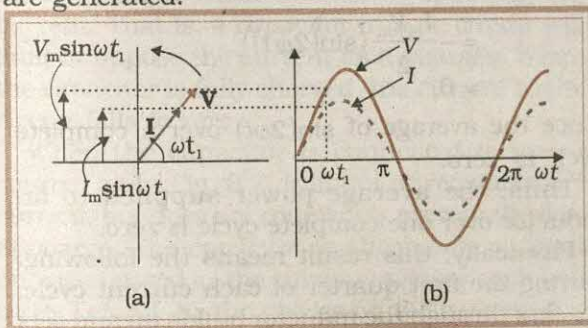


Fig. 8.4 (a) A phasor diagram for the circuit in Fig 8.1. (b) Graph of V and I versus ωt .

From Fig. 8.4(a) we see that phasors \mathbf{V} and \mathbf{I} for the case of a resistor are in the same direction. This is so for all times. This means that the phase angle between the voltage and the current is zero.

8.4 AC VOLTAGE APPLIED TO AN INDUCTOR

Figure 8.5 shows an ac source connected to an inductor. Usually, inductors have appreciable resistance in their windings, but we shall assume that this inductor has negligible resistance. Thus, the circuit is a purely inductive

ac circuit. Let the voltage across the source be $V = V_m \sin \omega t$. Using the Kirchhoff's loop rule, $\sum \epsilon(t) = 0$, and since there is no resistor in the circuit,

$$V - L \frac{dI}{dt} = 0 \quad (8.10)$$

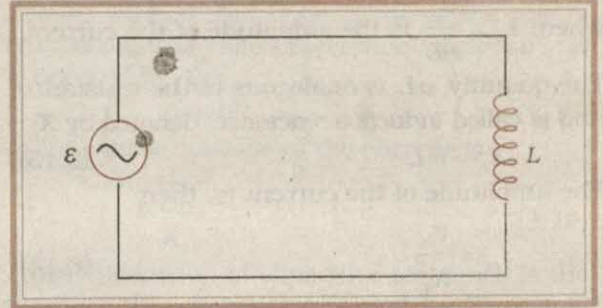


Fig. 8.5 An ac source connected to an inductor.

where the second term is the self-induced Faraday emf in the inductor; and L is the self-inductance of the inductor. The negative sign follows from Lenz's law (Chapter 7). Combining Eqs. (8.1) and (8.10), we have

$$\frac{dI}{dt} = \frac{V}{L} = \frac{V_m}{L} \sin \omega t \quad (8.11)$$

Equation (8.11) implies that the equation for $I(t)$, the current as a function of time, must be such that its slope dI/dt is a sinusoidally varying quantity, with the same phase as the source

voltage and an amplitude given by $\frac{V_m}{L}$. To obtain

the current, we integrate $\frac{dI}{dt}$ with respect to time:

$$\int \frac{dI}{dt} dt = \frac{V_m}{L} \int \sin(\omega t) dt$$

We get, $I = -\frac{V_m}{\omega L} \cos(\omega t) + \text{constant}$.

The integration constant has the dimension of current and is time-independent. Since the source has an emf which oscillates symmetrically about zero, the current it sustains also oscillates symmetrically about zero, so that no constant

* Though voltage and current in ac circuit are represented by phasors – rotating vectors, they are not vectors themselves. They are scalar quantities. It so happens that the amplitudes and phases of harmonically varying scalars combine mathematically in the same way as do the projections of rotating vectors of corresponding magnitudes and directions. The 'rotating vectors' that 'represent' harmonically varying scalar quantities are introduced only to provide us with a simple way of adding these quantities using a rule that we already know.

or time-independent component of the current exists. Therefore, the integration constant is zero. Using $-\cos(\omega t) = \sin\left(\omega t - \frac{\pi}{2}\right)$, we have

$$I = I_m \sin\left(\omega t - \frac{\pi}{2}\right) \quad (8.12)$$

where $I_m = \frac{V_m}{\omega L}$ is the amplitude of the current.

The quantity ωL is analogous to the resistance and is called *inductive reactance*, denoted by X_L :

$$X_L = \omega L \quad (8.13)$$

The amplitude of the current is, then

$$I_m = \frac{V_m}{X_L} \quad (8.14)$$

The dimension of inductive reactance is the same as that of resistance and its SI unit is ohm(Ω). The inductive reactance limits the current in a purely inductive circuit in the same way as the resistance limits the current in a purely resistive circuit. The inductive reactance is directly proportional to the inductance and to the frequency of the current.

A comparison of Eqs. (8.1) and (8.12) for the source voltage and the current in an inductor shows that the current lags the voltage by $\frac{\pi}{2}$ or one-quarter ($\frac{1}{4}$) cycle. Figure 8.6(a) shows the voltage and the current phasors in the present case at instant t_1 . The current phasor \mathbf{I} is $\frac{\pi}{2}$ behind the voltage phasor \mathbf{V} . When rotated with frequency ω counter-clockwise, they generate the voltage and current given by Eqs. (8.1) and (8.12), respectively and as shown in Fig. 8.6(b).

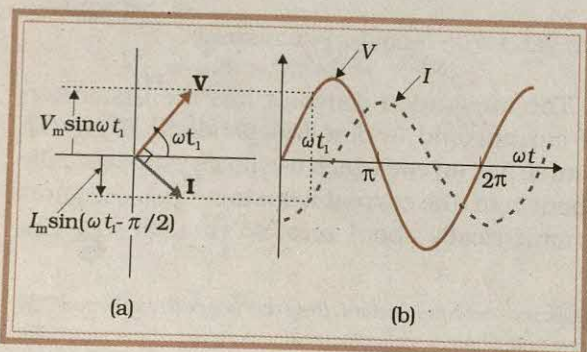


Fig. 8.6 (a) A Phasor diagram for the circuit in Fig. 8.5. (b) Graph of V and I versus ωt .

We see that the current reaches its maximum value later than the voltage by one-fourth of a period [$\frac{T}{4} = \frac{\pi/2}{\omega}$]. You have seen that an inductor has reactance that limits current similar to resistance in a dc circuit. Does it also consume power like a resistance? Let us try to find out.

The instantaneous power supplied to the inductor is

$$\begin{aligned} P_L &= IV = I_m \sin\left(\omega t - \frac{\pi}{2}\right) V_m \sin(\omega t) \\ &= -I_m V_m \cos(\omega t) \sin(\omega t) \\ &= -\frac{I_m V_m}{2} \sin(2\omega t) \end{aligned}$$

So, the average power over a complete cycle is

$$\begin{aligned} \overline{P_L} &= \left\langle -\frac{I_m V_m}{2} \sin(2\omega t) \right\rangle \\ &= -\frac{I_m V_m}{2} \langle \sin(2\omega t) \rangle \\ &= 0, \end{aligned}$$

since the average of $\sin(2\omega t)$ over a complete cycle is zero.

Thus, the average power supplied to an inductor over one complete cycle is zero.

Physically, this result means the following. During the first quarter of each current cycle, the flux through the inductor builds up and sets up a magnetic field and energy is stored in the inductor. In the next quarter of cycle, as the current decreases, the flux decreases and the stored energy is returned to the source. Thus, in each half cycle, the energy which is withdrawn from the source is returned to it without any dissipation of power.

Example 8.2 A pure inductor of 25.0 mH is connected to a source of 220 V. Find the inductive reactance and rms current in the circuit if the frequency of the source is 50 Hz.

Answer The inductive reactance,

$$\begin{aligned} X_L &= 2\pi \nu L = 2 \times 3.14 \times 50 \times 25 \times 10^{-3} \text{ W} \\ &= 7.85 \Omega \end{aligned}$$

The rms current in the circuit is

$$I_{rms} = \frac{V_{rms}}{X_L} = \frac{220 \text{ V}}{7.85 \Omega} = 28.03 \text{ A}$$

8.5 AC VOLTAGE APPLIED TO A CAPACITOR

Figure 8.7 shows an ac source ε connected to a capacitor only, a purely capacitive ac circuit.

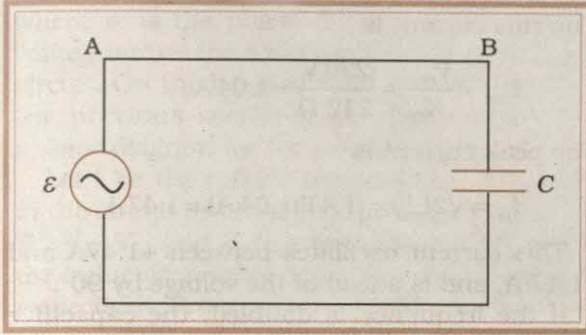


Fig. 8.7 An ac source connected to a capacitor.

When a capacitor is connected to a voltage source in a dc circuit, current will flow for the short time required to charge the capacitor. As charge accumulates on the capacitor plates, the voltage across them increases, opposing the current. That is, a capacitor in a dc circuit will limit or oppose the current as it charges. When the capacitor is fully charged, the current in the circuit falls to zero.

When the capacitor is connected to an ac source, as in Fig. 8.7, it limits or regulates the current, but does not completely prevent the flow of charge. The capacitor is alternately charged and discharged as the current reverses each half cycle. Let $q(t)$ be the charge on the capacitor at any time t . The instantaneous voltage $V(t)$ across the capacitor is

$$V(t) = \frac{q(t)}{C} \quad (8.15)$$

From the Kirchhoff's loop rule, the voltage across the source and the capacitor are equal,

$$V_m \sin \omega t = \frac{q}{C}$$

To find the current, we use the relation $I = \frac{dq}{dt}$

$$I = \frac{d}{dt}(V_m C \sin \omega t) = \omega C V_m \cos(\omega t)$$

Using the relation, $\cos(\omega t) = \sin\left(\omega t + \frac{\pi}{2}\right)$, we have

$$I = I_m \sin\left(\omega t + \frac{\pi}{2}\right) \quad (8.16)$$

where the amplitude of the oscillating current is $I_m = \omega C V_m$. We can rewrite it as

$$I_m = \frac{V_m}{(1/\omega C)}$$

Comparing it to $I_m = \frac{V_m}{R}$ for a purely resistive circuit, we find that $(1/\omega C)$ plays the role of resistance. It is called *capacitive reactance* and is denoted by X_c .

$$X_c = 1/\omega C \quad (8.17)$$

so that the amplitude of the current is

$$I_m = \frac{V_m}{X_c} \quad (8.18)$$

The dimension of capacitive reactance is the same as that of resistance and its SI unit is Ohm(Ω). The capacitive reactance limits the amplitude of the current in a purely capacitive circuit in the same way as the resistance limits the current in a purely resistive circuit. But it is inversely proportional to the frequency and the capacitance.

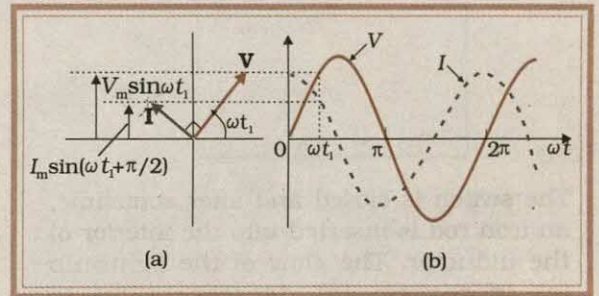


Fig. 8.8 (a) A Phasor diagram for the circuit in Fig. 8.7. (b) Graph of V and I versus ωt .

A comparison of Eq. (8.16) with the equation of source voltage, Eq. (8.1) shows that the current is $\frac{\pi}{2}$ rad ahead of voltage. Figure 8.8(a) shows the phasor diagram at an instant t_1 . Here the current phasor \mathbf{I} is $\frac{\pi}{2}$ rad ahead of the voltage phasor \mathbf{V} as they rotate counter clockwise. Figure 8.8(b) shows the variation of voltage and current with time. We see that the current reaches its maximum value earlier than the voltage by one-fourth of a period.

The instantaneous power supplied to the capacitor is

$$P_c = IV = I_m \cos(\omega t) \cdot V_m \sin(\omega t)$$

$$\begin{aligned}
 &= I_m V_m \cos(\omega t) \sin(\omega t) \\
 &= \frac{I_m V_m}{2} \sin(2\omega t) \quad (8.19)
 \end{aligned}$$

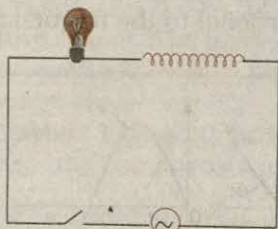
So, as in the case of an inductor, the average power

$$\bar{P}_C = \left\langle \frac{I_m V_m}{2} \sin(2\omega t) \right\rangle = \frac{I_m V_m}{2} \langle \sin(2\omega t) \rangle = 0$$

since $\langle \sin(2\omega t) \rangle = 0$ over a complete cycle. As discussed in the case of an inductor, the energy stored by a capacitor in each quarter period is returned to the source in the next quarter period.

Thus, we see that in the case of an inductor, the current lags the voltage by 90° and in the case of a capacitor, the current leads the voltage by 90° .

Example 8.3 A light bulb and an open coil inductor are connected to an ac source through a key as shown in the Figure.



The switch is closed and after sometime, an iron rod is inserted into the interior of the inductor. The glow of the lightbulb (a) increases; (b) decreases; (c) is unchanged, as the iron rod is inserted. Give your answer with reasons.

Answer As the iron rod is inserted, the magnetic field inside the coil is increased. Hence, the inductance of the coil increases. Consequently, the inductive reactance of the coil increases. As a result, a larger fraction of the applied ac voltage appears across the inductor, leaving less voltage across the bulb. Therefore, the glow of the light bulb decreases.

Example 8.4 A $15.0 \mu\text{F}$ capacitor is connected to a 220 V , 50 Hz source. Find the capacitive reactance and the current (rms and peak) in the circuit. If the frequency is doubled, what happens to the capacitive reactance and the current.

Answer The capacitive reactance is

$$X_C = \frac{1}{2\pi\nu C} = \frac{1}{2\pi(50\text{Hz})(15.0 \times 10^{-6}\text{F})} = 212 \Omega$$

The rms current is

$$I_{\text{rms}} = \frac{V_{\text{rms}}}{X_C} = \frac{220 \text{ V}}{212 \Omega} = 1.04 \text{ A}$$

The peak current is

$$I_m = \sqrt{2} I_{\text{rms}} = (1.41)(1.04 \text{ A}) = 1.47 \text{ A}$$

This current oscillates between $+1.47 \text{ A}$ and -1.47 A , and is ahead of the voltage by 90° .

If the frequency is doubled, the capacitive reactance is halved and consequently, the current is doubled.

8.6 AC VOLTAGE APPLIED TO A SERIES LCR CIRCUIT

Figure 8.9 shows a series LCR circuit connected to an ac source ε . As usual, we take the voltage of the source to be $V = V_m \sin \omega t$.

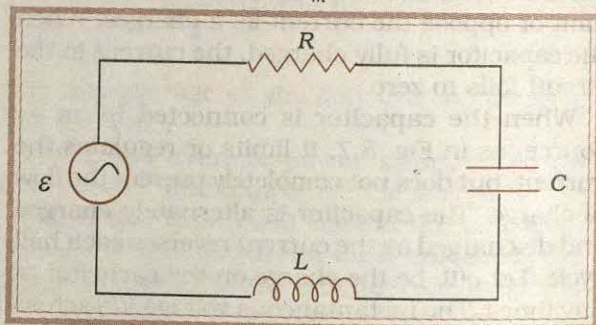


Fig. 8.9 A series LCR circuit connected to an ac source.

If q is the charge on the capacitor and I the current, at time t , we have, from Kirchhoff's loop rule:

$$L \frac{dI}{dt} + IR + \frac{q}{C} = V \quad (8.20)$$

We want to determine the instantaneous current I and its phase relationship to the applied alternating voltage V . We shall solve this problem by two methods. First, we use the technique of phasors and in the second method, we solve Eq. (8.20) analytically to obtain the time-dependence of I .

8.6.1 Phasor-diagram Solution

From the circuit shown in Fig. 8.9, we see that the resistor, inductor and capacitor are in series. Therefore, the ac current in each element is the

same, having the same amplitude and phase. Let it be

$$I = I_m \sin(\omega t + \phi) \quad (8.21)$$

where ϕ is the phase difference between the voltage across the source and the current in the circuit. On the basis of what we have learnt in the previous sections, we shall construct a phasor diagram for the present case.

Let \mathbf{I} be the phasor representing the current in the circuit as given by Eq. (8.21). Further, let \mathbf{V}_L , \mathbf{V}_R , \mathbf{V}_C , and \mathbf{V} represent the voltage across the inductor, resistor, capacitor and the source, respectively. From previous section, we know that

\mathbf{V}_R is parallel to \mathbf{I} , \mathbf{V}_C is $\frac{\pi}{2}$ rad behind \mathbf{I} and \mathbf{V}_L is

$\frac{\pi}{2}$ rad ahead of \mathbf{I} . \mathbf{V}_L , \mathbf{V}_R , \mathbf{V}_C and \mathbf{I} are shown in

Fig. 8.10(a) with appropriate phase-relations.

The length of these phasors or the amplitude of \mathbf{V}_R , \mathbf{V}_C and \mathbf{V}_L are:

$$V_{Rm} = I_m R, \quad V_{Cm} = I_m X_C, \quad V_{Lm} = I_m X_L \quad (8.22)$$

The voltage Equation (8.20) for the circuit can be written as

$$V_L + V_R + V_C = V \quad (8.23)$$

The phasor relation whose vertical component gives the above equation is

$$\mathbf{V}_L + \mathbf{V}_R + \mathbf{V}_C = \mathbf{V} \quad (8.24)$$

This relation is represented in Fig. 8.10(b). Since \mathbf{V}_C and \mathbf{V}_L are always along the same line and in opposite directions, they can be combined into a single phasor $(\mathbf{V}_C + \mathbf{V}_L)$ which has a magnitude $|V_{Cm} - V_{Lm}|$. Since, \mathbf{V} is represented as the hypotenuse of a right-angled triangle whose sides are \mathbf{V}_R and $(\mathbf{V}_C + \mathbf{V}_L)$, the pythagorean theorem gives:

$$V_m^2 = V_{Rm}^2 + (V_{Cm} - V_{Lm})^2$$

Substituting the values of V_{Rm} , V_{Cm} , and V_{Lm} from Eq. (8.22) into the above equation, we have

$$\begin{aligned} V_m^2 &= (I_m R)^2 + (I_m X_C - I_m X_L)^2 \\ &= I_m^2 [R^2 + (X_C - X_L)^2] \end{aligned}$$

$$\text{or} \quad I_m = \frac{V_m}{\sqrt{R^2 + (X_C - X_L)^2}} \quad (8.25a)$$

By analogy to the resistance in a circuit, we introduce the *impedance* Z in an ac circuit:

$$I_m = \frac{V_m}{Z} \quad (8.25b)$$

$$\text{where } Z = \sqrt{R^2 + (X_C - X_L)^2} \quad (8.26)$$

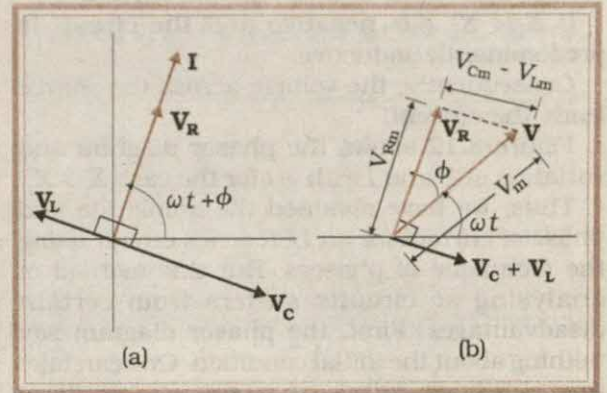


Fig. 8.10 (a) Relation between the phasors \mathbf{V}_L , \mathbf{V}_R , \mathbf{V}_C and \mathbf{I} , (b) Relation between the phasors \mathbf{V}_L , \mathbf{V}_R , and $(\mathbf{V}_L + \mathbf{V}_C)$ for the circuit in Fig. 8.9.

Since phasor \mathbf{I} is always parallel to phasor \mathbf{V}_R , the phase angle ϕ is the angle between \mathbf{V}_R and \mathbf{V} and can be determined from Fig. 8.11:

$$\tan \phi = \frac{V_{Cm} - V_{Lm}}{V_{Rm}}$$

Using Eq. (8.22), we have

$$\tan \phi = \frac{X_C - X_L}{R} \quad (8.27)$$

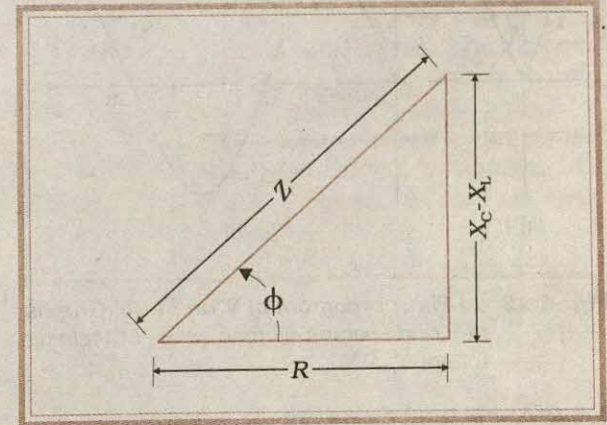


Fig. 8.11 Impedance diagram.

Equations (8.26) and (8.27) are graphically shown in Fig. (8.11). This is called *Impedance diagram* which is a right triangle with Z as its hypotenuse.

Equation 8.25(a) gives the amplitude of the current and Eq. (8.27) gives the phase angle. With these, Eq. (8.21) is completely specified.

If $X_C > X_L$, ϕ is positive and the circuit is predominantly capacitive. Consequently, the voltage across the source lags the current.

If $X_C < X_L$, ϕ is negative and the circuit is predominantly inductive.

Consequently, the voltage across the source leads the current.

Figure 8.12 shows the phasor diagram and variation of V and I with ωt for the case $X_C > X_L$.

Thus, we have obtained the amplitude and phase of current for an LCR series circuit using the technique of phasors. But this method of analysing ac circuits suffers from certain disadvantages. First, the phasor diagram says nothing about the initial condition. One can take any arbitrary value of t (say, t_1 , as done throughout in this Chapter) and draw different phasors which show the relative angle between different phasors. The solution so obtained is called the *steady-state solution*. This is not a general solution. Additionally, we do have a *transient solution* which exists even for $V = 0$. The general solution is the sum of the transient solution and the steady-state solution. After a sufficiently long time, the effects of the transient solution die out and the behaviour of the circuit is described by the steady-state solution.

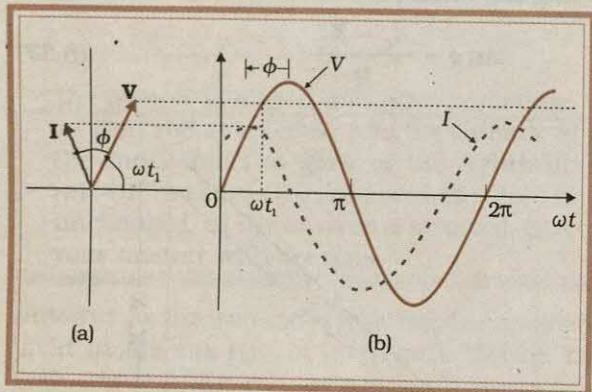


Fig. 8.12 (a) Phasor diagram of V and I . (b) Graphs of V and I versus ωt for a series LCR circuit where $X_C > X_L$.

8.6.2 Analytical Solution

The voltage equation for the circuit is

$$L \frac{dI}{dt} + RI + \frac{q}{C} = V = V_m \sin \omega t$$

We know that $I = \frac{dq}{dt}$. Therefore, $\frac{dI}{dt} = \frac{d^2q}{dt^2}$.

Thus, in terms of q , the voltage equation becomes

$$L \frac{d^2q}{dt^2} + R \frac{dq}{dt} + \frac{q}{C} = V_m \sin \omega t \quad (8.28)$$

Let us assume a solution

$$q = q_m \sin(\omega t + \theta) \quad (8.29a)$$

$$\text{so that } \frac{dq}{dt} = q_m \omega \cos(\omega t + \theta) \quad (8.29b)$$

$$\text{and } \frac{d^2q}{dt^2} = -q_m \omega^2 \sin(\omega t + \theta) \quad (8.29c)$$

Substituting these values in Eq. (8.28), we get

$$q_m \omega [R \cos(\omega t + \theta) + (X_C - X_L) \sin(\omega t + \theta)] = V_m \sin \omega t \quad (8.30)$$

where we have used the relation $X_C = \frac{1}{\omega C}$,

$X_L = \omega L$. Multiplying and dividing Eq. (8.30) by

$Z = \sqrt{R^2 + (X_C - X_L)^2}$, we have

$$q_m \omega Z \left[\frac{R}{Z} \cos(\omega t + \theta) + \frac{(X_C - X_L)}{Z} \sin(\omega t + \theta) \right] = V_m \sin \omega t \quad (8.31)$$

Now, let $\frac{R}{Z} = \cos \phi$

and $\frac{(X_C - X_L)}{Z} = \sin \phi$

so that $\phi = \tan^{-1} \frac{X_C - X_L}{R} \quad (8.32)$

Substituting this in Eq. (8.31) and simplifying, we get:

$$q_m \omega Z \cos(\omega t + \theta - \phi) = V_m \sin \omega t \quad (8.33)$$

Comparing the two sides of this equation, we see that

$$V_m = q_m \omega Z = I_m Z$$

where

$$I_m = q_m \omega \quad (8.33a)$$

$$\text{and } \theta - \phi = -\frac{\pi}{2} \quad \text{or } \theta = -\frac{\pi}{2} + \phi \quad (8.33b)$$

Therefore, the current in the circuit is

$$I = \frac{dq}{dt} = q_m \omega \cos(\omega t + \theta) = I_m \cos(\omega t + \theta)$$

$$\text{or } I = I_m \sin(\omega t + \phi) \quad (8.34)$$

$$\text{where } I_m = \frac{V_m}{Z} = \frac{V_m}{\sqrt{R^2 + (X_C - X_L)^2}} \quad 8.34(a)$$

$$\text{and } \phi = \tan^{-1} \frac{X_C - X_L}{R}$$

Thus, the analytical solution for the amplitude and phase of the current in the circuit agrees with that obtained by the technique of phasors.

8.6.3 Resonance

An interesting characteristic of the series RLC circuit is the phenomenon of resonance. The phenomenon of resonance is common among systems that have a tendency to oscillate at a particular frequency. This frequency is called the system's natural frequency. If such a system is driven by an energy source at a frequency that is near the natural frequency, the amplitude of oscillation is found to be large. A familiar example of this phenomenon is a child on a swing. The child on the swing has a natural frequency for swinging back and forth. If the child pulls on the rope at regular intervals and the frequency of the pulls is almost the same as the frequency of swinging, the amplitude of the swinging will be large (Chapter 14, Class XII).

For an RLC circuit driven with voltage of amplitude V_m and frequency ω , we found that the current amplitude is given by

$$I_m = \frac{V_m}{Z} = \frac{V_m}{\sqrt{R^2 + (X_C - X_L)^2}}$$

with $X_C = \frac{1}{\omega C}$ and $X_L = \omega L$. So if ω is varied,

then at a particular frequency ω_0 $X_C = X_L$, and the impedance is minimum ($Z = \sqrt{R^2 + 0^2} = R$). This frequency is called the *resonant frequency*:

$$X_C = X_L \text{ or } \frac{1}{\omega_0 C} = \omega_0 L$$

$$\text{or } \omega_0 = \frac{1}{\sqrt{LC}} \quad (8.35)$$

At resonant frequency, the current amplitude is maximum; $I_m = \frac{V_m}{R}$

Figure 8.13 shows the variation of I_m with ω in a RLC series circuit with $L = 1.00$ mH,

$C = 1.00$ nF for two values of R : (i) $R = 100 \Omega$ and (ii) $R = 200 \Omega$. For the source applied $V_m = 100$ V, ω_0 for this case is $(\frac{1}{\sqrt{LC}}) = 1.00 \times 10^6$ rad/s.

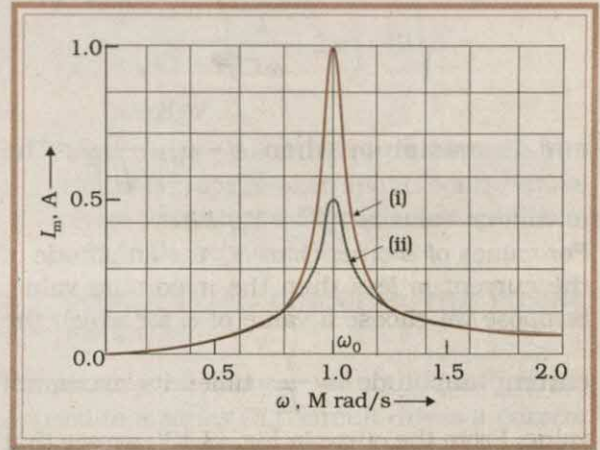


Fig. 8.13 Variation of I_m with ω for two cases: (i) $R = 100 \Omega$, (ii) $R = 200 \Omega$. $L = 1.00$ mH, $C = 1.00$ nF and $V_m = 100$ V for both cases.

We see that the current amplitude is maximum

at the resonant frequency. Since $I_m = \frac{V_m}{R}$ at resonance, the current amplitude for case (i) is twice to that for case (ii).

Resonant circuits have a variety of applications, for example, in the tuning mechanism of a radio or a TV set. The antenna of a radio accepts signals from many broadcasting stations. The antenna acts as a source in the tuning circuit of the radio, so the circuit can be driven at many frequencies. But to hear one particular radio station, we tune the radio. In tuning, we vary the capacitance of a capacitor in the tuning circuit such that the resonant frequency of the circuit becomes nearly equal to the frequency of the radio signal received. When this happens, the amplitude of the current in the circuit is maximum.

It is important to note that resonance phenomenon is exhibited by a circuit only if both L and C are present in the circuit. Only then do the voltages across L and C cancel each other (both being 180° out of phase) and the current

amplitude is $\frac{V_m}{R}$, the total source voltage

appearing across R . This means that we can not have resonance in a RL or RC circuit.

Sharpness of resonance

The amplitude of the current in the series LCR circuit is given by

$$I_m = \frac{V_m}{\sqrt{R^2 + \left(\omega L - \frac{1}{\omega C}\right)^2}}$$

and is maximum when $\omega = \omega_0 = \frac{1}{\sqrt{LC}}$. The

maximum value is $I_m^{\max} = V_m / R$.

For values of ω other than ω_0 , the amplitude of the current is less than the maximum value. Suppose we choose a value of ω for which the

current amplitude is $\frac{1}{\sqrt{2}}$ times its maximum

value. From the curve in Fig. (8.13), we see that there are two such values of ω , say, ω_1 and ω_2 , one greater and the other smaller than ω_0 and symmetrical about ω_0 . We may write

$$\omega_1 = \omega_0 + \Delta\omega$$

$$\omega_2 = \omega_0 - \Delta\omega$$

The difference $\omega_1 - \omega_2 = 2\Delta\omega$ is often called the *bandwidth* of the circuit. The quantity $(\omega_0 / 2\Delta\omega)$ is regarded as a measure of the sharpness of resonance. The smaller the $\Delta\omega$, the sharper or narrower is the resonance.

To get an expression for $\Delta\omega$, we note that the current amplitude I_m is $\frac{1}{\sqrt{2}} I_m^{\max}$ for $\omega_1 = \omega_0 + \Delta\omega$. Therefore,

$$\begin{aligned} \text{at } \omega_1 \quad I_m &= \frac{V_m}{\sqrt{R^2 + \left(\omega_1 L - \frac{1}{\omega_1 C}\right)^2}} \\ &= \frac{I_m^{\max}}{\sqrt{2}} = \frac{V_m}{R\sqrt{2}} \end{aligned}$$

$$\text{or} \quad \sqrt{R^2 + \left(\omega_1 L - \frac{1}{\omega_1 C}\right)^2} = R\sqrt{2}$$

$$\text{or} \quad R^2 + \left(\omega_1 L - \frac{1}{\omega_1 C}\right)^2 = 2R^2$$

$$\omega_1 L - \frac{1}{\omega_1 C} = R$$

which may be written as,

$$(\omega_0 + \Delta\omega)L - \frac{1}{(\omega_0 + \Delta\omega)C} = R$$

$$\omega_0 L \left(1 + \frac{\Delta\omega}{\omega_0}\right) - \frac{1}{\omega_0 C \left(1 + \frac{\Delta\omega}{\omega_0}\right)} = R$$

Using $\omega_0^2 = \frac{1}{LC}$ in the second term on the left hand side, we get

$$\omega_0 L \left(1 + \frac{\Delta\omega}{\omega_0}\right) - \frac{\omega_0 L}{\left(1 + \frac{\Delta\omega}{\omega_0}\right)} = R$$

We can approximate $\left(1 + \frac{\Delta\omega}{\omega_0}\right)^{-1}$ as $\left(1 - \frac{\Delta\omega}{\omega_0}\right)$ since

$\frac{\Delta\omega}{\omega_0} \ll 1$. Therefore,

$$\omega_0 L \left(1 + \frac{\Delta\omega}{\omega_0}\right) - \omega_0 L \left(1 - \frac{\Delta\omega}{\omega_0}\right) = R$$

$$\text{or} \quad \omega_0 L \frac{2\Delta\omega}{\omega_0} = R$$

$$\Delta\omega = \frac{R}{2L} \quad 8.36(a)$$

The sharpness of resonance is given by,

$$\frac{\omega_0}{2\Delta\omega} = \frac{\omega_0 L}{R} \quad 8.36(b)$$

The ratio $\frac{\omega_0 L}{R}$ is also called the **quality factor**, Q of the circuit.

$$Q = \frac{\omega_0 L}{R} \quad 8.36(c)$$

From Eqs. [8.36(b)] and [8.36(c)], we see that

$2\Delta\omega = \frac{\omega_0}{Q}$. So, larger the value of Q , the smaller is the value of $2\Delta\omega$ or the bandwidth and sharper

is the resonance. Using $\omega_0^2 = \frac{1}{LC}$, Eq. [8.36(c)]

can be equivalently expressed as $Q = \frac{1}{\omega_0 CR}$.

We see from Fig. 8.13, that if the resonance is less sharp, not only is the maximum current less, the circuit is close to resonance for a larger range $\Delta\omega$ of frequencies and the tuning of the circuit will not be good. So, less sharp the resonance, less is the selectivity of the circuit or vice-versa. From Eq., (8.36), we see that if quality factor is large i.e., R is low or L is large, the circuit is more selective.

Example 8.5 A resistor of $200\ \Omega$ and a capacitor of $15.0\ \mu\text{F}$ are connected in series to a $220\ \text{V}$, $50\ \text{Hz}$ ac source. (a) Calculate the current in the circuit; (b) Calculate the voltage (rms) across the resistor and the capacitor. Is the algebraic sum of these voltages more than the source voltage? If yes, resolve the paradox.

Answer Given

$$R = 200\ \Omega, C = 15.0\ \mu\text{F} = 15.0 \times 10^{-6}\ \text{F}$$

$$V_{\text{rms}} = 220\ \text{V}, \nu = 50\ \text{Hz}$$

- (a) In order to calculate the current, we need the impedance of the circuit. It is

$$\begin{aligned} Z &= \sqrt{R^2 + X_C^2} = \sqrt{R^2 + (2\pi \nu C)^{-2}} \\ &= \sqrt{(200\ \Omega)^2 + (2 \times 3.14 \times 50 \times 10^{-6}\ \text{F})^{-2}} \\ &= \sqrt{(200\ \Omega)^2 + (212\ \Omega)^2} \\ &= 291.5\ \Omega \end{aligned}$$

Therefore, the current in the circuit is

$$I_{\text{rms}} = \frac{V_{\text{rms}}}{Z} = \frac{220\ \text{V}}{291.5\ \Omega} = 0.755\ \text{A}$$

- (b) Since the current is the same throughout the circuit, we have

$$V_R = I_{\text{rms}} R = (0.755\ \text{A})(200\ \Omega) = 151\ \text{V}$$

$$V_C = I_{\text{rms}} X_C = (0.755\ \text{A})(212.3\ \Omega) = 160.3\ \text{V}$$

The algebraic sum of the two voltages, V_R and V_C is $311.3\ \text{V}$ which is more than the source voltage of $220\ \text{V}$. How to resolve this paradox? As you have learnt in the text, the

two voltages are not in the same phase. Therefore, they cannot be added like ordinary numbers. The two voltages are out of phase by ninety degrees. Therefore, the total of these voltages must be obtained using the Pythagorean theorem:

$$\begin{aligned} V_{R+C} &= \sqrt{V_R^2 + V_C^2} \\ &= 220\ \text{V} \end{aligned}$$

Thus, if the phase difference between two voltages is properly taken into account, the total voltage across the resistor and the capacitor is equal to the voltage of the source.

8.7 POWER IN AC CIRCUITS: THE POWER FACTOR

We have seen that a voltage $V = V_m \sin \omega t$ applied to a series RLC circuit drives a current in the circuit given by $I = I_m \sin(\omega t + \phi)$ where

$$I_m = \frac{V_m}{Z} \text{ and } \phi = \tan^{-1} \left(\frac{X_C - X_L}{R} \right)$$

Therefore, the instantaneous power P supplied to the source is

$$\begin{aligned} P &= VI = (V_m \sin \omega t) \times [I_m \sin(\omega t + \phi)] \\ &= \frac{V_m I_m}{2} [\cos \phi - \cos(2\omega t + \phi)] \end{aligned} \quad (8.37)$$

The average power over a cycle is given by the average of the two terms in R.H.S. of Eq. (8.37). It is only the second term which is time-dependent. Its average is zero (the positive half of the cosine cancels the negative half). Therefore,

$$\begin{aligned} \bar{P} &= \frac{V_m I_m}{2} \cos \phi \\ &= \frac{V_m}{\sqrt{2}} \frac{I_m}{\sqrt{2}} \cos \phi \\ &= V_{\text{rms}} I_{\text{rms}} \cos \phi \end{aligned} \quad (8.38\text{a})$$

This can also be written as,

$$\bar{P} = I_{\text{rms}}^2 Z \cos \phi \quad (8.38\text{b})$$

So, the average power dissipated depends not only on the voltage and current but also on the cosine of the phase angle ϕ between them. The quantity $\cos \phi$ is called the *power factor*. Let us discuss the following cases:

Case(i) Resistive circuit: If the circuit contains only pure R , it is called *resistive*. In that case $\phi = 0$, $\cos \phi = 1$. There is maximum power dissipation.

Case(ii) Purely inductive or capacitive circuit: If the circuit contains only an inductor or capacitor, we know that the phase difference

between voltage and current is $\frac{\pi}{2}$. Therefore,

$\cos \phi = 0$, and no power is dissipated even though a current is flowing in the circuit. This current is sometimes referred to as *wattless current*.

Case(iii) LCR series circuit: In an LCR series circuit, power dissipated is given by Eq. (8.38)

where $\phi = \tan^{-1} \frac{X_C - X_L}{R}$. So, ϕ may be non-zero

(except $\frac{\pi}{2}$) in a RL or RC or RCL circuit. Even in such cases, power is dissipated only in the resistor.

Case(iv) Power dissipated at resonance in LCR circuit: At resonance $X_C - X_L = 0$, and $\phi = 0$.

Therefore, $\cos \phi = 1$, and $P = \frac{I_m}{Z} \cdot \frac{I_m}{R}$. That is, maximum power is dissipated in a circuit (through R) at resonance.

Example 8.6 A sinusoidal voltage of peak value 283 V and frequency 50 Hz is applied to a series LCR circuit in which $R = 3 \Omega$, $L = 25.48 \text{ mH}$, and $C = 796 \mu\text{F}$. Find (a) the impedance of the circuit; (b) the phase difference between the voltage across the source and the current; (c) the power dissipated in the circuit; and (d) the power factor.

Answer

(a) To find the impedance of the circuit, we first calculate X_L and X_C .

$$X_L = 2\pi\nu L$$

$$= 2 \times 3.14 \times 50 \times 25.48 \times 10^{-3} \Omega = 8 \Omega$$

$$X_C = \frac{1}{2\pi\nu C}$$

$$= \frac{1}{2 \times 3.14 \times 50 \times 796 \times 10^{-6}} = 4 \Omega$$

Therefore,

$$Z = \sqrt{R^2 + (X_L - X_C)^2} = \sqrt{3^2 + (8 - 4)^2} = 5 \Omega$$

$$(b) \text{ Phase difference, } \phi = \tan^{-1} \frac{X_C - X_L}{R}$$

$$= \tan^{-1} \left(\frac{4 - 8}{3} \right) = -53.1^\circ$$

since ϕ is negative, the current in the circuit lags the voltage across the source.

(c) The power dissipated in the circuit is

$$P = I_{rms}^2 R$$

$$\text{Now, } I_{rms} = \frac{I_m}{\sqrt{2}} = \frac{1}{\sqrt{2}} \left(\frac{283}{5} \right) = 40 \text{ A}$$

$$\text{Therefore, } P = (40 \text{ A})^2 \times 3 \Omega = 4800 \text{ W}$$

$$(d) \text{ Power factor} = \cos \phi = \cos 53.1^\circ = 0.6.$$

Example 8.7 Suppose the frequency of the source in the previous example can be varied. (a) What is the frequency of the source at which resonance occurs? (b) Calculate the impedance, the current, and the power dissipated at the resonant condition.

Answer

(a) The frequency at which the resonance occurs is

$$\omega_0 = \frac{1}{\sqrt{LC}} = \frac{1}{\sqrt{25.48 \times 10^{-3} \times 796 \times 10^{-6}}} = 222.1 \text{ rad/s}$$

$$\nu_r = \frac{\omega_0}{2\pi} = \frac{222.1}{2 \times 3.14} \text{ Hz} = 35.4 \text{ Hz}$$

(b) The impedance Z at resonant condition is equal to the resistance:

$$Z = R = 3 \Omega$$

The rms current at resonance is

$$= \frac{V_{rms}}{Z} = \frac{V_{rms}}{R} = \left(\frac{283}{\sqrt{2}} \right) \frac{1}{3} = 66.7 \text{ A}$$

The power dissipated at resonance is

$$P = I_{rms}^2 \times R = (66.7)^2 \times 3 = 13.35 \text{ kW}$$

You can see that in the present case, power dissipated at resonance is more than the power dissipated in Example 8.6.

Example 8.8 At an airport, a person is made to walk through the doorway of a metal detector, for security reasons. If she/he is carrying anything made of metal, the metal detector emits a sound. On what principle does this detector work?

Answer The metal detector works on the principle of resonance in ac circuits. When you walk through a metal detector, you are, in fact, walking through a coil of many turns. The coil is connected to a capacitor tuned so that the circuit is in resonance. When you walk through with metal in your pocket, the impedance of the circuit changes – resulting in significant change in current in the circuit. This change in current is detected and the electronic circuitry causes a sound to be emitted as an alarm.

8.8 LC OSCILLATIONS

We know that a capacitor and an inductor can store electrical and magnetic energy, respectively. When a capacitor (initially charged) is connected to an inductor, the charge on the capacitor and the current in the circuit exhibit the phenomenon of electrical oscillations similar to oscillations in mechanical systems (Chapter 14, Class XI).

Let a capacitor be charged q_0 (at $t = 0$) and connected to an inductor as shown in Fig. 8.14.

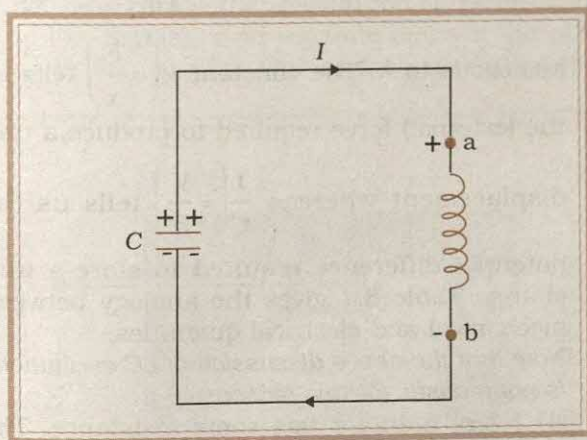


Fig. 8.14 At the instant shown, the current is increasing so the polarity of induced emf in the inductor is as shown.

The moment the circuit is completed, the charge on the capacitor starts decreasing, giving rise to current in the circuit. Let q and I be the

charge and current in the circuit at time t . Since $\frac{dI}{dt}$ is positive, the induced emf in L will have polarity as shown, i.e., $V_b < V_a$. According to Kirchhoff's loop rule,

$$\frac{q}{C} - L \frac{dI}{dt} = 0 \quad (8.39)$$

$I = -\frac{dq}{dt}$ in the present case (as q decreases, I increases). Therefore, Eq. (8.39) becomes:

$$\frac{d^2q}{dt^2} + \frac{1}{LC}q = 0 \quad (8.40)$$

This equation has the form $\frac{d^2x}{dt^2} + \omega_0^2 x = 0$ for a simple harmonic oscillator. The charge, therefore, oscillates with a natural frequency

$$\omega_0 = \frac{1}{\sqrt{LC}} \quad (8.41)$$

and varies sinusoidally with time as

$$q = q_m \cos(\omega_0 t + \phi) \quad (8.42)$$

where q_m is the maximum value of q and ϕ is a phase constant. Since $q = q_m$ at $t = 0$, we have $\cos \phi = 1$ or $\phi = 0$. Therefore, in the present case,

$$q = q_m \cos(\omega_0 t) \quad (8.43)$$

The current $I \left(= -\frac{dq}{dt} \right)$ is given by

$$I = I_m \sin(\omega_0 t) \quad (8.44)$$

where $I_m = \omega_0 q_m$.

Let us now try to visualise how this oscillation takes place in the circuit.

Figure 8.15(a) shows a capacitor with initial charge q_m connected to an ideal inductor. The electrical energy stored in the charged capacitor

is $U_E = \frac{1}{2} \frac{q_m^2}{C}$. Since, there is no current in the circuit, energy in the inductor is zero. Thus, the total energy of LC circuit is,

$$U = U_E = \frac{1}{2} \frac{q_m^2}{C}$$

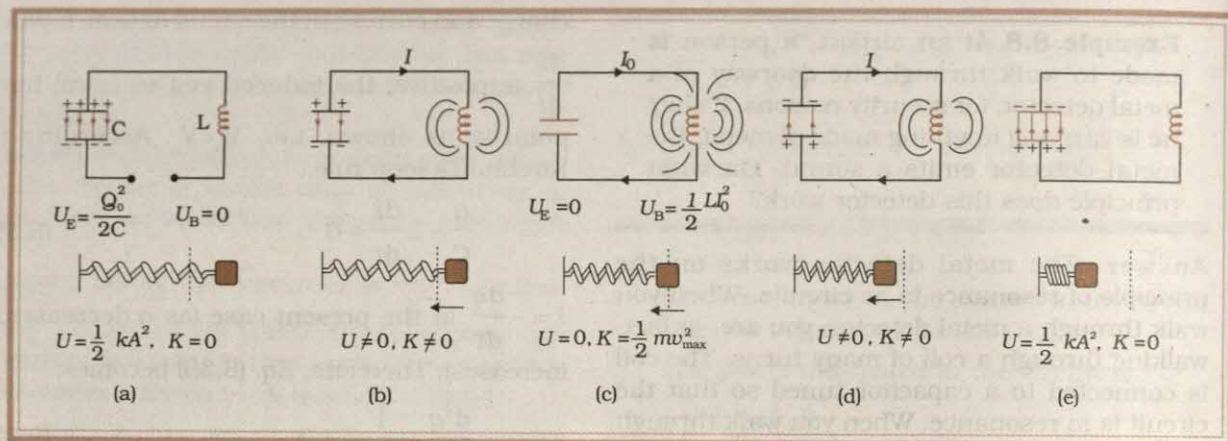


Fig. 8.15 The oscillations in an LC circuit are analogous to the oscillation of a block at the end of a spring. The figure depicts one-half of a cycle.

At $t = 0$, the switch is closed and the capacitor starts to discharge [Fig. 8.15(b)]. As the current increases, it sets up a magnetic field in the inductor and thereby, some energy gets stored in the inductor in the form of magnetic energy:

$U_B = \frac{1}{2} LI^2$. As the current reaches its maximum value I_0 , as in Fig. 8.15(c), all the energy is stored

in the magnetic field: $U_B = \frac{1}{2} LI_0^2$. You can easily

check that the maximum electrical energy equals the maximum magnetic energy. The capacitor now has no charge and hence no energy. The current now starts charging the capacitor, as in Fig. 8.15(d). This process continues till the capacitor is fully charged [Fig. 8.15(e)]. But it is charged with a polarity opposite to its initial state in Fig. 8.15(a). The whole process just described will now repeat itself till the system reverts to its original state. Thus, the energy in the system oscillates between the capacitor and the inductor.

The LC oscillation is similar to the mechanical oscillation of a block attached to a spring. The lower part of each figure in Fig. 8.16 depicts the corresponding stage of a mechanical system (a block attached to a spring). As noted earlier, for a block of a mass m oscillating with frequency ω_0 , the equation is

$$\frac{d^2x}{dt^2} + \omega_0^2 x = 0$$

Here, $\omega_0 = \sqrt{\frac{k}{m}}$, and k is the spring constant.

So, x corresponds to q . In case of a mechanical

system $F = ma = m \frac{dv}{dt} = m \frac{d^2x}{dt^2}$. For an electrical

system, $\varepsilon = -L \frac{dI}{dt} = -L \frac{d^2q}{dt^2}$. Comparing these

two equations, we see that L is analogous to mass m : L is a measure of resistance to change

in current. In case of LC circuit, $\omega_0 = \frac{1}{\sqrt{LC}}$ and

for mass on a spring, $\omega_0 = \sqrt{\frac{k}{m}}$. So, $\frac{1}{C}$ is

analogous to k . The constant $k \left(= \frac{F}{x} \right)$ tells us

the (external) force required to produce a unit

displacement whereas $\frac{1}{C} \left(= \frac{V}{q} \right)$ tells us the

potential difference required to store a unit charge. Table 8.1 gives the analogy between mechanical and electrical quantities.

Note that the above discussion of LC oscillations is not realistic for two reasons:

- (1) Every inductor has some resistance. The effect of this resistance is to introduce a damping effect on the charge and current in the circuit and the oscillations finally die away.
- (2) Even if the resistance were zero, the total energy of the system would not remain constant. It is radiated away from the system

Table 8.1 Analogies between Mechanical and Electrical Quantities

Mechanical System	Electrical System
Mass m	Inductance L
Force constant k	Reciprocal capacitance $1/C$
Displacement x	Charge q
Velocity $v = dx/dt$	Current $I = dq/dt$
Mechanical energy	Electromagnetic energy
$E = \frac{1}{2}kx^2 + \frac{1}{2}mv^2$	$U = \frac{1}{2} \frac{q^2}{C} + \frac{1}{2}LI^2$

in the form of electromagnetic waves (discussed in the next Chapter). In fact, radio and TV transmitters depend on this radiation.

8.9 TRANSFORMERS

For many purposes, it is necessary to change (or transform) an alternating voltage from one to another of greater or smaller value. This is done with a device called *transformer* using the principle of mutual induction.

A transformer consists of two sets of coils, insulated from each other. They are wound on a soft-iron core, either one on top of the other as in Fig. 8.16(a) or on separate limbs of the core as in Fig. 8.16(b). One of the coils called the

primary coil has N_p turns. The other coil is called the secondary coil; it has N_s turns. Often the primary coil is the input coil and the secondary coil is the output coil of the transformer.

When an alternating voltage is applied to the primary, the resulting current produces an alternating magnetic flux which links the secondary and induces an emf in it. The value of this emf depends on the number of turns in the secondary. We consider an ideal transformer in which the primary has negligible resistance and all the flux in the core links both primary and secondary windings. Let ϕ be the flux in each turn in the core at time t due to current in the primary when a voltage V_p is applied to it. Then the induced emf or voltage ε_s , in the secondary with N_s turns is

$$\varepsilon_s = -N_s \frac{d\phi}{dt} \quad (8.45)$$

The alternating flux ϕ also induces an emf, called back emf in the primary. This is

$$\varepsilon_p = -N_p \frac{d\phi}{dt} \quad (8.46)$$

But $\varepsilon_p = V_p$. If this were not so, the primary current would be infinite since the primary has zero resistance (as assumed). If the secondary is an open circuit or the current taken from it is small, then to a good approximation

$$\varepsilon_s = V_s$$

where V_s is the voltage across the secondary. Therefore, Eqs. (8.45) and (8.46) can be written as

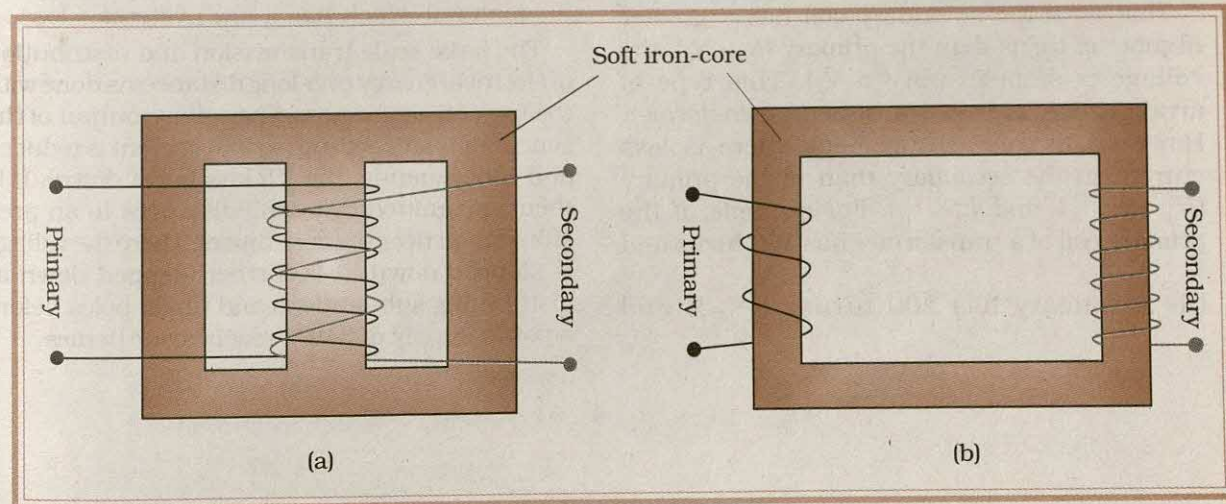


Fig. 8.16 Two arrangements for winding of primary and secondary coil in a transformer: (a) two coils on top of each other, (b) two coils on separate limbs of the core.

$$V_s = -N_s \frac{d\phi}{dt} \quad [8.45(a)]$$

$$V_p = -N_p \frac{d\phi}{dt} \quad [8.46(a)]$$

From Eqs. [8.45(a)] and [8.46(a)], we have

$$\frac{V_s}{V_p} = \frac{N_s}{N_p} \quad (8.47)$$

Note that the above relation has been obtained using three assumptions: (i) the primary resistance and current are small; (ii) the same flux links both the primary and the secondary as very little flux escapes from the core, and (iii) the secondary current is small.

If the transformer is assumed to be 100% efficient (no energy losses), the power input is equal to the power output, and since $P = IV$,

$$I_p V_p = I_s V_s \quad (8.48)$$

Although some energy is always lost, this is a good approximation, since a well designed transformer may have an efficiency of more than 95%. Combining Eqs. (8.47) and (8.48), we have

$$\frac{I_p}{I_s} = \frac{V_s}{V_p} = \frac{N_s}{N_p} \quad (8.49)$$

Now, we can see how a transformer affects the voltage and current. We have:

$$V_s = \left(\frac{N_s}{N_p} \right) V_p \quad \text{and} \quad I_s = \left(\frac{N_p}{N_s} \right) I_p \quad (8.50)$$

That is, if the secondary coil has a greater number of turns than the primary ($N_s > N_p$), the voltage is stepped up ($V_s > V_p$). This type of arrangement is called a *step-up transformer*. However, in this arrangement, there is less current in the secondary than in the primary ($N_p / N_s < 1$ and $I_s < I_p$). For example, if the primary coil of a transformer has 100 turns and

the secondary has 200 turns, $\frac{N_s}{N_p} = 2$ and

$\frac{N_p}{N_s} = \frac{1}{2}$. Thus, a 220 V input at 10 A will step-up to 440 V output at 5.0 A.

If the secondary coil has less turns than the primary ($N_s < N_p$), we have a *step-down transformer*. In this case, $V_s < V_p$ and $I_s > I_p$. That is, the voltage is stepped down, or reduced, and the current is increased.

The equations obtained above apply to ideal transformers (without any energy losses). But in actual transformers, small energy losses do occur due to the following reasons:

- (i) *Flux Leakage*: There is always some flux leakage; that is, not all of the flux due to primary passes through the secondary due to poor design of the core or the air gaps in the core. It can be reduced by winding the primary and secondary coils one over the other.
- (ii) *Resistance of the windings*: The wire used for the windings has some resistance and so, energy is lost due to heat produced in the wire ($I^2 R$). In high current, low voltage windings, these are minimised by using thick wire.
- (iii) *Eddy currents*: The alternating magnetic flux induces eddy currents in the iron core and causes heating. The effect is reduced by having a laminated core.
- (iv) *Hysteresis*: The magnetisation of the core is repeatedly reversed by the alternating magnetic field. The resulting expenditure of energy in the core appears as heat and is kept to a minimum by using a magnetic material which has a low hysteresis loss.

The large scale transmission and distribution of electrical energy over long distances is done with the use of transformers. The voltage output of the generator is stepped-up (so that current is reduced and consequently, the $I^2 R$ loss is cut down). It is then transmitted over long distances to an area sub-station near the consumers. There the voltage is stepped down. It is further stepped down at distributing sub-stations and utility poles before a power supply of 240 V reaches our homes.

SUMMARY

1. An alternating voltage $V = V_m \sin \omega t$ applied to a resistor R drives a current $I = I_m \sin \omega t$ in the resistor, $I_m = V_m/R$. The current is in phase with the applied voltage.
2. For an alternating current $I = I_m \sin \omega t$ passing through a resistor R , the average power loss \bar{P} (averaged over a cycle) due to joule heating is $(1/2)I_m^2 R$. To express it in the same form as the dc power ($P = I^2 R$), a special value of current is used. It is called *root mean square (rms) current* and is denoted by I_{rms} :

$$I_{rms} = \frac{I_m}{\sqrt{2}} = 0.707 I_m$$

Similarly, the *rms voltage* is defined by

$$V_{rms} = \frac{V_m}{\sqrt{2}} = 0.707 V_m$$

We have $\bar{P} = I_{rms} V_{rms} = I_{rms}^2 R$.

3. An ac voltage $V = V_m \sin \omega t$ applied to a pure inductor L , drives a current in the inductor $I = I_m \sin(\omega t - \pi/2)$, where $I_m = \frac{V_m}{X_L}$, $X_L = \omega L$ is called *inductive reactance*. The current in the inductor lags the voltage by $\pi/2$. The average power supplied to an inductor over one complete cycle is zero.
4. An ac voltage $V = V_m \sin \omega t$ applied to a capacitor drives a current in the capacitor:

$$I = I_m \sin(\omega t + \pi/2). \text{ Here, } I_m = \frac{V_m}{X_C}, X_C = \frac{1}{\omega C} \text{ is called } \textit{capacitive reactance}.$$

The current through the capacitor is $\pi/2$ ahead of the applied voltage. As in the case of inductor, the average power supplied to a capacitor over one complete cycle is zero.

5. For a series *RLC* circuit driven by voltage $V = V_m \sin \omega t$, the current is given by

$$I = I_m \sin(\omega t + \phi)$$

$$\text{where } I_m = \frac{V_m}{\sqrt{R^2 + (X_C - X_L)^2}}$$

$$\text{and } \phi = \tan^{-1} \frac{X_C - X_L}{R}.$$

$$Z = \sqrt{R^2 + (X_C - X_L)^2} \text{ is called the } \textit{impedance} \text{ of the circuit.}$$

The average power loss over a complete cycle is given by

$$\bar{P} = V_{rms} I_{rms} \cos \phi$$

The term $\cos \phi$ is called the *power factor*.

6. In a purely inductive or capacitive circuit, $\cos \phi = 0$ and no power is dissipated even though a current is flowing in the circuit. In such cases, current is referred to as a *wattless current*.
7. The phase relationship between current and voltage in an ac circuit can be shown conveniently by representing voltage and current by rotating vectors called *phasors*. A phasor is a vector which rotates about the origin with angular speed ω . The magnitude of a phasor represents the amplitude or peak value of the quantity (voltage or current) represented by the phasor.

The analysis of an ac circuit is facilitated by the use of a phasor diagram.

8. An interesting characteristic of a series RLC circuit is the phenomenon of *resonance*. The circuit exhibits resonance, i.e., the amplitude of the current is maximum at the resonant frequency, $\omega_0 = \frac{1}{\sqrt{LC}}$. The *quality*

factor Q defined by $Q = \frac{\omega_0 L}{R} = \frac{1}{\omega_0 CR}$ is an indicator of the sharpness of the resonance, the higher value of Q indicating sharper peak in the current.

9. A circuit containing an inductor L and a capacitor C (initially charged) with no ac source and no resistors exhibits *free oscillations*. The charge q of the capacitor satisfies the equation of simple harmonic motion:

$$\frac{d^2 q}{dt^2} + \frac{1}{LC} q = 0$$

and therefore, the frequency ω of free oscillation is $\omega_0 = \frac{1}{\sqrt{LC}}$. The energy

in the system oscillates between the capacitor and the inductor but their sum or the total energy is constant in time.

10. A transformer consists of an iron core on which are bound a primary coil of N_p turns and a secondary coil of N_s turns. If the primary coil is connected to an ac source, the primary and secondary voltages are related by

$$V_s = \left(\frac{N_s}{N_p} \right) V_p$$

and the currents are related by

$$I_s = \left(\frac{N_p}{N_s} \right) I_p.$$

If the secondary coil has a greater number of turns than the primary, the voltage is stepped-up ($V_s > V_p$). This type of arrangement is called a *step-up transformer*.

If the secondary coil has turns less than the primary, we have a *step-down transformer*.

Physical quantity	Symbol	Dimensions	Unit	Remarks
rms voltage	V_{rms}	$[M L^2 T^{-3} A^{-1}]$	V	$V_{rms} = \frac{V_m}{\sqrt{2}}$, V_m is the amplitude of the ac voltage.
rms current	I_{rms}	[A]	A	$I_{rms} = \frac{I_m}{\sqrt{2}}$, I_m is the amplitude of the ac current.
Reactance:				
Inductive	X_L	$[M L^2 T^{-3} A^{-2}]$	Ω	$X_L = \omega L$
Capacitive	X_C	$[M L^2 T^{-3} A^{-2}]$	Ω	$X_C = 1/\omega C$
Impedance	Z	$[M L^2 T^{-3} A^{-2}]$	Ω	Depends on elements present in the circuit.
Resonant frequency	ω_r or ω_0	$[T^{-1}]$	Hz	$\omega_r = \frac{1}{\sqrt{LC}}$ for a series RLC circuit
Quality factor	Q	Dimensionless		$Q = \frac{\omega_r L}{R} = \frac{1}{\omega_r C R}$ for a series RLC circuit.
Power factor		Dimensionless		$= \cos \phi$, ϕ is the phase difference between voltage applied and current in the circuit.

POINTS TO PONDER

- When a value is given for ac voltage or current, it is ordinarily the rms value. The voltage across the terminals of an outlet in your room is normally 240 V. This refers to the rms value of the voltage. The amplitude of this voltage is

$$V_m = \sqrt{2} V_{rms} = \sqrt{2}(240) = 340 \text{ V}$$

- The power rating of an element used in ac circuits refers to its average power rating.
- Though the average current over a cycle in an ac circuit is zero, the average power is not zero.
- Both alternating current and direct current are measured in amperes. But how is the ampere physically defined for an alternating current? It cannot be derived from the mutual attraction of two parallel wires carrying ac

currents as the dc ampere is derived. An ac current changes direction with the source frequency and the attractive force would average to zero. Thus, the ac ampere must be defined in terms of some property that is independent of the direction of the current. Joule heating is such a property, and there is one ampere of *rms* value of alternating current in a circuit if the current produces the same average heating effect as one ampere of dc current would produce under the same conditions.

5. In an ac circuit, while adding voltages across different elements, one should take care of their phases properly. For example, if V_R and V_C are voltages across R and C , respectively in an RC circuit, then the total voltage across RC combination is $V_{RC} = \sqrt{V_R^2 + V_C^2}$ and not $V_R + V_C$ since V_C is $\frac{\pi}{2}$ out of phase of V_R .
6. Though in a phasor diagram, voltage and current are represented by vectors, these quantities are not really vectors themselves. They are scalar quantities. It so happens that the amplitudes and phases of harmonically varying scalars combine mathematically in the same way as do the projections of rotating vectors of corresponding magnitudes and directions. The 'rotating vectors' that represent harmonically varying scalar quantities are introduced only to provide us with a simple way of adding these quantities using a rule that we already know as the law of vector addition.
7. There are no power losses associated with pure capacitances and pure inductances in an ac circuit. The only element that dissipates energy in an ac circuit is the resistive element.
8. In a RLC circuit, resonance phenomenon occurs when $X_L = X_C$ or $\omega_0 = \frac{1}{\sqrt{LC}}$. For resonance to occur, the presence of both L and C elements in the circuit is a must. With only one of these (L or C) elements, there is no possibility of voltage cancellation and hence, no resonance is possible.
9. The power factor in a RLC circuit is a measure of how close the circuit is to expending the maximum power.
10. In generators and motors, the roles of input and output are reversed. In a motor, electric energy is the input and mechanical energy is the output. In a generator, mechanical energy is the input and electric energy is the output. Both devices simply transform energy from one form to another.
11. A transformer (step-up) changes a low-voltage into a high-voltage. This does not violate the law of conservation of energy. The current is reduced by the same proportion.
12. The choice of whether the description of an oscillatory motion is by means of sines or cosines or by their linear combinations is unimportant, since changing the zero-time position transforms the one to the other.

EXERCISES

- 8.1** A $100\ \Omega$ resistor is connected to a 220 V, 50 Hz ac supply.
 (a) What is the rms value of current in the circuit?
 (b) What is the net power consumed over a full cycle?
- 8.2** (a) The peak voltage of an ac supply is 300 V. What is the rms voltage?
 (b) The rms value of current in an ac circuit is 10 A. What is the peak current?
- 8.3** A 44 mH inductor is connected to 220 V, 50 Hz ac supply. Determine the rms value of the current in the circuit.
- 8.4** A $60\ \mu\text{F}$ capacitor is connected to a 110 V, 60 Hz ac supply. Determine the rms value of the current in the circuit.
- 8.5** In Exercises 8.3 and 8.4, what is the net power absorbed by each circuit over a complete cycle. Explain your answer.
- 8.6** Show that a series LCR circuit driven by an ac source exhibits resonance at $\omega_r = 1/\sqrt{LC}$.
- 8.7** Obtain the resonant frequency ω_r of a series LCR circuit with $L = 2.0\text{H}$, $C = 32\ \mu\text{F}$ and $R = 10\ \Omega$. What is the Q -value of this circuit?
- 8.8** Why is a choke coil needed in the use of fluorescent tubes with ac mains? Why can we not use an ordinary resistor instead of the choke coil?
- 8.9** Show that the angular frequency of free oscillations of an LC circuit is equal to $1/\sqrt{LC}$.
- 8.10** Show that in the free oscillations of an LC circuit, the sum of energies stored in the capacitor and the inductor is constant in time.
- 8.11** A charged $30\ \mu\text{F}$ capacitor is connected to a 27 mH inductor. What is the angular frequency of free oscillations of the circuit?
- 8.12** Suppose the initial charge on the capacitor in Exercise 8.11 is 6 mC. What is the total energy stored in the circuit initially? What is the total energy at later time?
- 8.13** A series LCR circuit with $R = 20\ \Omega$, $L = 1.5\text{H}$ and $C = 35\ \mu\text{F}$ is connected to a variable-frequency 200 V ac supply. When the frequency of the supply equals the natural frequency of the circuit, what is the average power transferred to the circuit in one complete cycle?
- 8.14** A radio can tune over the frequency range of a portion of MW broadcast band: (800 kHz to 1200 kHz). If its LC circuit has an effective inductance of $200\ \mu\text{H}$, what must be the range of its variable capacitor?
[Hint: For tuning, the natural frequency i.e., the frequency of free oscillations of the LC circuit should be equal to the frequency of the radiowave.]
- 8.15** Figure 8.17 shows a series LCR circuit connected to a variable frequency 230 V source. $L = 5.0\text{H}$, $C = 80\ \mu\text{F}$, $R = 40\ \Omega$.
 (a) Determine the source frequency which drives the circuit in resonance.
 (b) Obtain the impedance of the circuit and the amplitude of current at the resonating frequency.
 (c) Determine the rms potential drops across the three elements of the circuit. Show that the potential drop across the LC combination is zero at the resonating frequency.

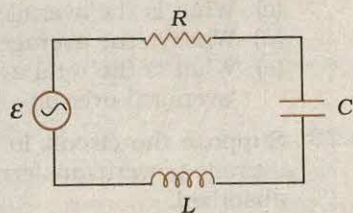


Fig. 8.17

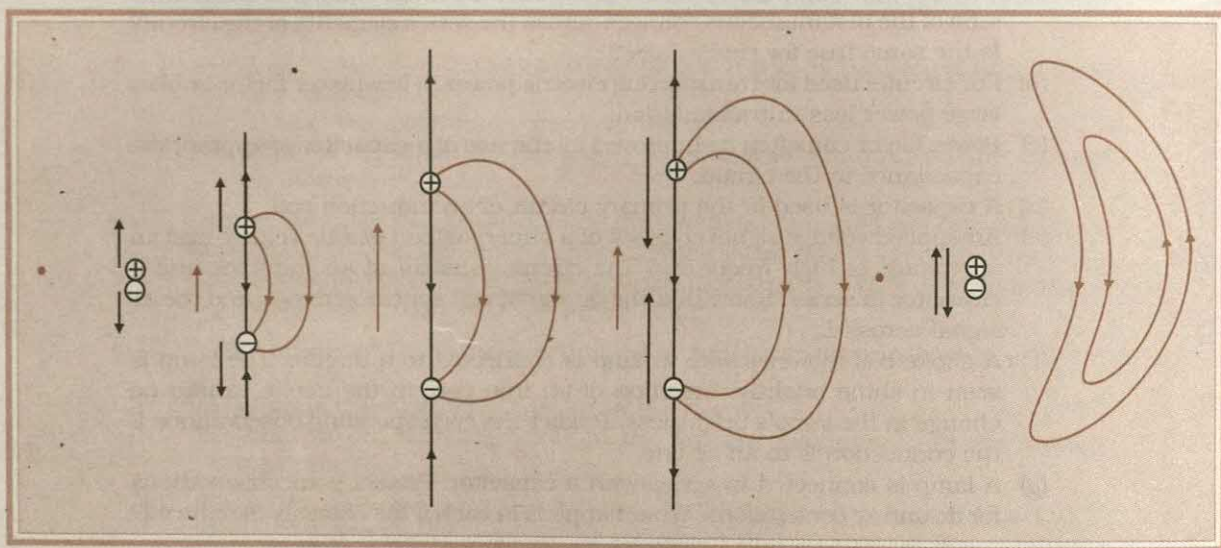
ADDITIONAL EXERCISES

- 8.16** An LC circuit contains a 20 mH inductor and a $50\text{ }\mu\text{F}$ capacitor with an initial charge of 10 mC . The resistance of the circuit is negligible. Let the instant the circuit is closed be $t = 0$.
- What is the total energy stored initially? Is it conserved during LC oscillations?
 - What is the natural frequency of the circuit?
 - At what time is the energy stored
 - completely electrical (i.e., stored in the capacitor)?
 - completely magnetic (i.e., stored in the inductor)?
 - At what times is the total energy shared equally between the inductor and the capacitor?
 - If a resistor is inserted in the circuit, how much energy is eventually dissipated as heat?
- 8.17** A coil of inductance 0.50 H and resistance $100\text{ }\Omega$ is connected to a 240 V , 50 Hz ac supply.
- What is the maximum current in the coil?
 - What is the time lag between the voltage maximum and the current maximum?
- 8.18** Obtain the answers (a) to (b) above if the circuit is connected to a high frequency supply (240 V , 10 kHz). Hence, explain the statement that at very high frequency, an inductor in a circuit nearly amounts to an open circuit. How does an inductor behave in a dc circuit after the steady state?
- 8.19** A $100\text{ }\mu\text{F}$ capacitor in series with a $40\text{ }\Omega$ resistance is connected to a 110 V , 60 Hz supply.
- What is the maximum current in the circuit?
 - What is the time lag between the current maximum and the voltage maximum?
- 8.20** Obtain the answers to (a) and (b) above if the circuit is connected to a 110 V , 12 kHz supply? Hence, explain the statement that a capacitor is a conductor at very high frequencies. Compare this behaviour with that of a capacitor in a dc circuit after the steady state.
- 8.21** Keeping the source frequency equal to the resonating frequency of the series LCR circuit, if the three elements, L , C and R are arranged in parallel, show that the total current in the parallel LCR circuit is minimum at this frequency. Obtain the current rms value in each branch of the circuit for the elements and source specified in Exercise 8.15 for this frequency.
- 8.22** A circuit containing a 80 mH inductor and a $60\text{ }\mu\text{F}$ capacitor in series is connected to a 230 V , 50 Hz supply. The resistance of the circuit is negligible.
- Obtain the current amplitude and rms values.
 - Obtain the rms values of potential drops across each element.
 - What is the average power transferred to the inductor?
 - What is the average power transferred to the capacitor?
 - What is the total average power absorbed by the circuit? [Average implies 'averaged over one cycle'.]
- 8.23** Suppose the circuit in Exercise 8.22 has a resistance of $15\text{ }\Omega$. Obtain the average power transferred to each element of the circuit, and the total power absorbed.
- 8.24** A series LCR circuit with $L = 0.12\text{ H}$, $C = 480\text{ nF}$, $R = 23\text{ }\Omega$ is connected to a 230 V variable frequency supply.
- What is the source frequency for which current amplitude is maximum. Obtain this maximum value.

- (b) What is the source frequency for which average power absorbed by the circuit is maximum. Obtain the value of this maximum power.
- (c) For which frequencies of the source is the power transferred to the circuit half the power at resonant frequency? What is the current amplitude at these frequencies?
- (d) What is the Q -factor of the given circuit?
- 8.25** Obtain the resonant frequency and Q -factor of a series LCR circuit with $L = 3.0 \text{ H}$, $C = 27 \mu\text{F}$, and $R = 7.4 \Omega$. It is desired to improve the sharpness of the resonance of the circuit by reducing its 'full width at half maximum' by a factor of 2. Suggest a suitable way.
- 8.26** Answer the following questions:
- In any ac circuit, is the applied instantaneous voltage equal to the algebraic sum of the instantaneous voltages across the series elements of the circuit? Is the same true for rms voltage?
 - For circuits used for transporting electric power, a low power factor implies large power loss in transmission.
 - Power factor can often be improved by the use of a capacitor of appropriate capacitance in the circuit.
 - A capacitor is used in the primary circuit of an induction coil.
 - An applied voltage signal consists of a superposition of a dc voltage and an ac voltage of high frequency. The circuit consists of an inductor and a capacitor in series. Show that the dc signal will appear across C and the ac signal across L .
 - A choke coil in series with a lamp is connected to a dc line. The lamp is seen to shine brightly. Insertion of an iron core in the choke causes no change in the lamp's brightness. Predict the corresponding observations if the connection is to an ac line.
 - A lamp is connected in series with a capacitor. Predict your observations for dc and ac connections. What happens in each if the capacity is reduced?
- 8.27** A power transmission line feeds input power at 2300 V to a step-down transformer with its primary windings having 4000 turns. What should be the number of turns in the secondary in order to get output power at 230 V?
- 8.28** At a hydroelectric power plant, the water pressure head is at a height of 300 m and the water flow available is $100 \text{ m}^3 \text{ s}^{-1}$. If the turbine generator efficiency is 60%, estimate the electric power available from the plant ($g = 9.8 \text{ ms}^{-2}$).
- 8.29** A small town with a demand of 800 kW of electric power at 220 V is situated 15 km away from an electric plant generating power at 440 V. The resistance of the two wire line carrying power is 0.5Ω per km. The town gets power from the line through a 4000-220 V step-down transformer at a sub-station in the town.
- Estimate the line power loss in the form of heat.
 - How much power must the plant supply, assuming there is negligible power loss due to leakage?
 - Characterise the step up transformer at the plant.
- 8.30** Do the same exercise as above with the replacement of the earlier transformer by a 40,000-220 V step-down transformer (Neglect, as before, leakage losses though this may not be a good assumption any longer because of the very high voltage transmission involved). Hence, explain why high voltage transmission is preferred?

CHAPTER NINE

ELECTROMAGNETIC WAVES



9.1 INTRODUCTION

In Chapter 5, we learnt that an electric current produces magnetic field and that two current-carrying wires exert a magnetic force on each other. Further, in Chapter 7, we have seen that a magnetic field changing with time gives rise to an electric field. Is the converse also true? Does an electric field changing with time give rise to a magnetic field? James Clerk Maxwell (1831-1879), argued that this was indeed the case—not only an electric current but also a time-varying electric field generates magnetic field. He formulated a set of equations involving electric and magnetic fields, and their sources, the charge and current densities. These equations, that you will study in higher courses in detail are known as Maxwell's equations. Together with the Lorentz force formula (Chapter 5), they mathematically express all the basic laws of electromagnetism.

The most important prediction to emerge from Maxwell's equations is the existence of electromagnetic waves, which are (coupled) time-varying electric and magnetic fields propagating in space. The speed of the waves, according to his equations, turned out to be very close to the speed of light (3×10^8 m/s), obtained from optical measurements. This led to the remarkable conclusion that light is an electromagnetic wave. Maxwell's work thus unified the domain of electricity, magnetism and light. Hertz, in 1885, experimentally demonstrated the existence of electromagnetic waves. Its technological use by Marconi and others led in due course to the revolution in communication that we are witnessing today.



Heinrich Rudolf Hertz

(1857-1894)

German physicist who was the first to broadcast and receive radio waves. He produced electromagnetic waves, sent them through space, and measured their wavelength and speed. He showed that the nature of their vibration, reflection and refraction was the same as that of light and heat waves, establishing their identity for the first time. He also pioneered research on discharge of electricity through gases, and discovered the photoelectric effect.



James Clerk Maxwell

(1831-1879)

Born in Edinburgh, Scotland, was among the greatest physicists of the nineteenth century. He derived the thermal velocity distribution of molecules in a gas and was among the first to obtain reliable estimates of molecular parameters from measurable quantities like viscosity, etc. Maxwell's greatest achievement was the unification of the laws of electricity and magnetism (discovered by Coulomb, Oersted, Ampere and Faraday) into a consistent set of equations now called Maxwell's equations. From these he arrived at the most important conclusion that light is an electromagnetic wave. Interestingly, Maxwell did not agree with the idea (strongly suggested by the Faraday's laws of electrolysis) that electricity was particulate in nature.

9.2 ELECTROMAGNETIC WAVES

9.2.1 Sources of Electromagnetic Waves

How are electromagnetic waves produced? Neither stationary charges nor charges in uniform motion (steady currents) can be sources of electromagnetic waves. The former produces only electrostatic fields, while the latter also produces magnetic fields that, however, do not vary with time. It is an important result of Maxwell's theory that accelerated charges radiate electromagnetic waves. The proof of this basic result is beyond the scope of this book, but we can accept it on the basis of rough, qualitative reasoning. Consider a charge oscillating with some frequency. (An oscillating charge is an example of accelerating charge.) This produces an oscillating electric field in space, which produces an oscillating magnetic field, which in turn, is a source of oscillating electric field, and so on. The oscillating electric and magnetic fields thus regenerate each other, so to speak, as the wave propagates through the space. The frequency of the electromagnetic wave naturally equals the frequency of oscillation of the charge. The energy associated with the propagating wave comes at the expense of the energy of the source – the accelerated charge. In Section 9.2.3, we describe this non-mathematical picture of the origin of electromagnetic waves in some detail.

From the preceding discussion, it might appear easy to test the prediction that light is an electromagnetic wave. We might think that all we needed to do was to set up an ac circuit in which the current oscillated at the frequency of visible light, say, yellow light. But, alas, that is not possible. The frequency of yellow light is about 6×10^{14} Hz, while the frequency that we get even with modern electronic circuits is hardly above 10^{11} Hz. This is why the experimental demonstration of electromagnetic wave had to come in the low frequency region (the radio wave region), as in the Hertz's experiment that we describe next.

9.2.2 Hertz's Demonstration of Electromagnetic Waves

When Maxwell's work was published in 1867, it did not receive immediate acceptance. Many physicists were skeptical about the existence of electromagnetic waves. The experiment that conclusively demonstrated the existence of

electromagnetic waves was first performed by Heinrich Hertz in 1887. But Maxwell did not live to see this validation of his theory.

Figure 9.1 shows the experimental arrangement used by Hertz. Two large metal spheres S and S' are attached to two large metal plates P and P' respectively. The spheres are connected to an induction coil I . By interrupting currents in the induction coils, a sudden high voltage is applied across the gap. The voltage is high enough to ionise the air in the gap and a spark jumps the gap. Since the air is ionised, the spark gap consists of electrons and ions from the air, which oscillate back and forth. This process results in the production of electromagnetic waves. The frequency of electromagnetic waves is determined by the inductance and capacitance of the coils or rods that form the gap.

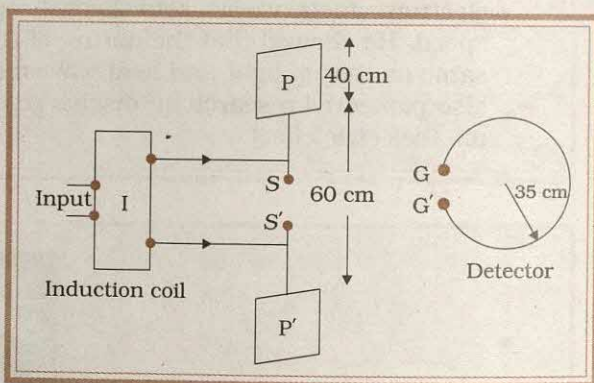


Fig. 9.1 Experimental arrangement used by Hertz for the production and detection of electromagnetic waves.

To detect these waves, Hertz designed a detector which consisted of a single loop of wire connected to two spheres. It had its own effective inductance, capacitance and natural frequency of oscillation. The electromagnetic waves reaching the gap of the detector had an electric field strong enough to establish a high potential difference between the gap GG' and cause a spark. Hertz could observe tiny sparks jumping between the gap GG' – thereby proving detection of electromagnetic radiations. Hertz also found that when the detector gap is at right angles to the source gap, no electromagnetic radiation is detected. Production of sparks between G and G' is maximum when the two gaps are parallel. This means that the electric vector of radiation produced by the source gap is parallel to the two gaps i.e., in a direction perpendicular to the

direction of propagation of the radiation. This clearly demonstrates that the electromagnetic (em) waves are transverse waves.

Hertz not only showed the existence of em waves, but also demonstrated that the waves, which had wavelength ten million times that of the light waves, could be diffracted, refracted and polarised. Thus, he conclusively established the wave nature of the radiation. Further, he produced stationary electromagnetic waves and determined their wavelength by measuring the distance between two successive nodes. Since the frequency of the wave was known (being equal to the frequency of the oscillator), he obtained the speed of the wave using the formula $v = \nu\lambda$ and found that the waves travelled with the same speed as the speed of light.

The fact that electromagnetic waves are polarised can be easily seen in the response of a portable AM radio to a broadcasting station. If an AM radio has a telescopic antenna, it responds to the electric part of the signal. When the antenna is turned horizontal, the signal will be greatly diminished. Some portable radios have horizontal antenna (usually inside the case of radio) which are sensitive to the magnetic component of the electromagnetic wave. Such a radio must remain horizontal in order to receive the signal. In such cases, response also depends on the orientation of the radio with respect to the station.

Hertz's successful experimental test of Maxwell's theory created a sensation and sparked off other important works in this field. Two important achievements in this connection deserve mention. Seven years after Hertz, Jagdish Chandra Bose, working at Calcutta, succeeded in producing and observing electromagnetic waves of much shorter wave length (25 mm to 5 mm). His experiment, like Hertz's, was confined to the laboratory.

At around the same time, Guglielmo Marconi in Italy followed Hertz's work and succeeded in transmitting electromagnetic waves over distances of many kilometres. Marconi's experiment marks the beginning of the field of communication using electromagnetic waves.

9.2.3 Qualitative Picture of Generation of Electromagnetic Waves

We can visualise the generation and propagation of electromagnetic waves in a qualitative way

using the concept of field lines developed by Faraday, as described in Chapter 1.

Consider a dipole antenna connected to an alternating voltage generator as shown in Fig. 9.2. This is a common technique for accelerating charged particles and is the source of radio waves emitted by the broadcast antenna of a radio station.

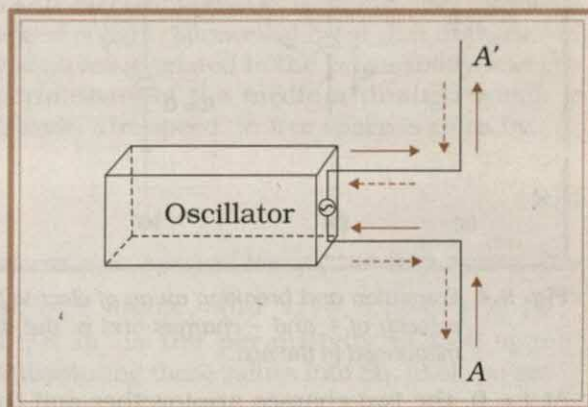


Fig. 9.2 A simple radio antenna that produces dipole radiation.

The charges accelerate in one direction (\rightarrow), then in the other direction (\leftarrow) alternatively in the arms of the antenna AA' . This is like an electric dipole oscillating with the frequency of the applied ac voltage.

You know that the dipole moment \mathbf{p} is equal to the product of either charge and the separation between the two charges. It is a vector directed from the negative to the positive charge. Since the separation between the two charges is varying with time, the magnitude and direction of dipole moment varies as shown in Fig. 9.3.

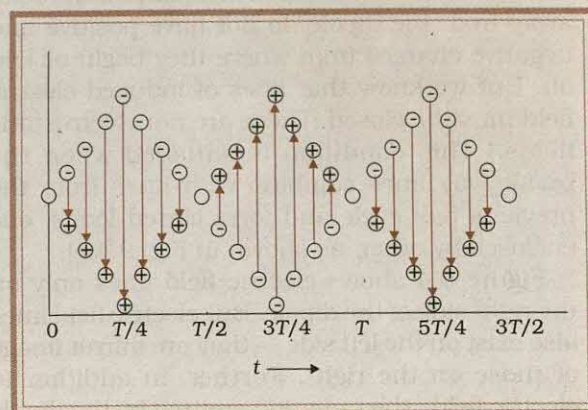


Fig. 9.3 Positions of the charges representing an oscillating electric dipole as a function of time.

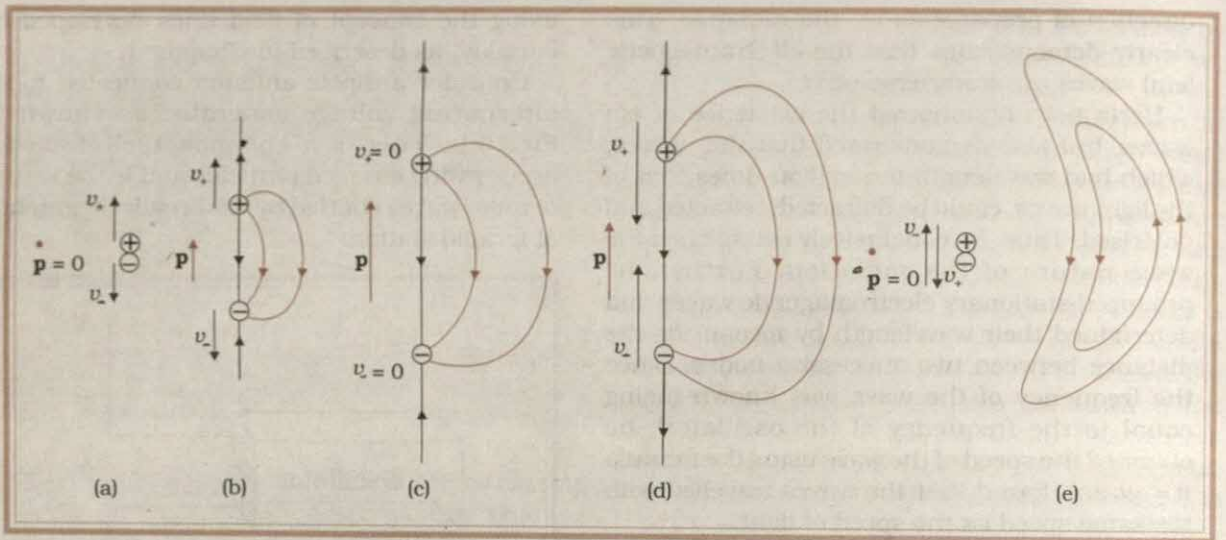


Fig. 9.4 Formation and breaking away of electric field lines from an oscillating dipole. v_+ and v_- represent the velocity of + and - charges and \mathbf{p} , the dipole moment of the system of two charges at the instant mentioned in the text.

At $t = 0$, the two charges are together and the distance between them is zero. Hence, the dipole moment is also zero, and there is no electric field [Fig. 9.4(a)].

When the separation between charges increases, the dipole moment increases from zero and the electric field lines start from the positive charge and end on the negative charge as shown in Fig. 9.4(b). The separation becomes maximum at $t = T/4$, the dipole moment is maximum and field lines are as shown in Fig. 9.4(c). After $t = T/4$, the charges start moving in the opposite direction and the dipole moment starts decreasing [Fig. 9.4(d)]. At $t = T/2$, the dipole moment is again zero. The electric field lines break away from the dipole. The lines that break away from the dipole do not have positive and negative charges from where they begin or end on. But we know that lines of induced electric field must be closed. (These are not electrostatic fields.) This condition is satisfied when the breakaway lines combine with lines from the previous half-cycle and form closed loops, one enclosed by other, as shown in Fig. 9.4(e).

Figure 9.4 shows electric field lines only on the right side of the dipole. But electric field lines also exist on the left side — they are mirror image of those on the right. Further, in addition to electric field, the radiation emitted by the dipole also has a magnetic field. The accelerating charges in the antenna produce a varying

current which produces a magnetic field. From the right-hand rule, the magnetic field will be perpendicular to the page, that is along the z -axis. This magnetic field at any point oscillates with the same frequency as that of the electric field. Combined together, they constitute the electromagnetic wave that propagates along the x -axis.

Figure 9.5 shows a complete picture of electric and magnetic field lines at an instant.

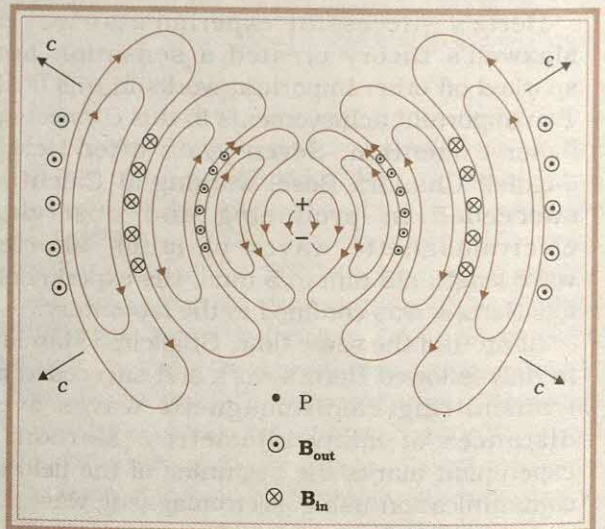


Fig. 9.5 Instantaneous pattern of the electric and magnetic field lines due to an oscillating dipole. The \mathbf{E} fields are shown by lines and the \mathbf{B} fields by dots and crosses.

To summarise, an oscillating electric dipole radiates electromagnetic waves. The wave is transverse with electric and magnetic fields perpendicular to each other and to the direction of propagation. If \mathbf{E} is along y -axis and the \mathbf{B} is along the z -axis, the direction of propagation is along $\mathbf{E} \times \mathbf{B}$, i.e., along the x -axis as shown in Fig. 9.6.

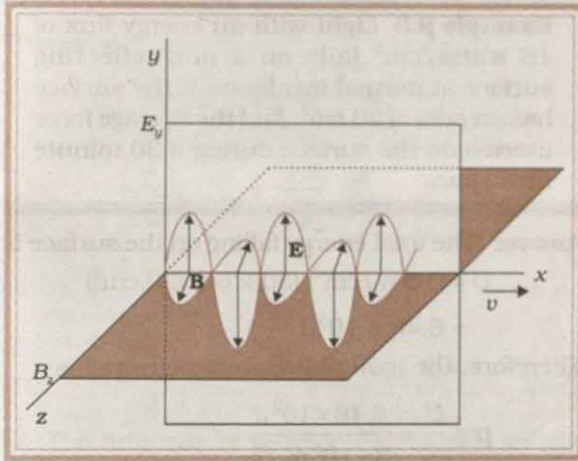


Fig. 9.6 Electromagnetic wave propagating along the x -axis. Electric and magnetic fields are along the y -axis and z -axis, respectively.

9.2.4 Nature of Electromagnetic Waves

The electric and magnetic fields shown in Fig. 9.6. are mathematically represented by:

$$\begin{aligned}\mathbf{E} &= E_y \hat{\mathbf{j}} = E_0 \sin [kx - \omega t] \hat{\mathbf{j}} \\ &= E_0 \sin \left[2\pi \left(\frac{x}{\lambda} - \nu t \right) \right] \hat{\mathbf{j}} \\ &= E_0 \sin \left[2\pi \left(\frac{x}{\lambda} - \frac{t}{T} \right) \right] \hat{\mathbf{j}}\end{aligned}\quad (9.1)$$

$$E_x = E_z = 0$$

$$\begin{aligned}\mathbf{B} &= B_z \hat{\mathbf{k}} = B_0 \sin [kx - \omega t] \hat{\mathbf{k}} \\ &= B_0 \sin \left[2\pi \left(\frac{x}{\lambda} - \nu t \right) \right] \hat{\mathbf{k}} \\ &= B_0 \sin \left[2\pi \left(\frac{x}{\lambda} - \frac{t}{T} \right) \right] \hat{\mathbf{k}}\end{aligned}\quad (9.2)$$

$$B_x = B_y = 0$$

where E_0 and B_0 are the amplitudes of the electric field \mathbf{E} and magnetic field \mathbf{B} , respectively.

$k = \frac{2\pi}{\lambda}$ is the magnitude of the wave vector \mathbf{k} . Other symbols have their usual meaning. The magnitude of \mathbf{E} and \mathbf{B} are related by,

$$\frac{E}{B} = c \text{ or } \frac{E_0}{B_0} = c \quad (9.3)$$

(The proof of this relation appears in higher courses.)

All electromagnetic waves travel with the speed of light. Maxwell showed that the speed of em waves is related to the permeability and the permittivity of the medium through which it travels. The speed in free space is given by,

$$c = \frac{1}{\sqrt{\mu_0 \epsilon_0}} \quad (9.4)$$

where $\mu_0 = 4\pi \times 10^{-7} \text{ N s}^2 / \text{C}^2$ is the permeability of free space, and $\epsilon_0 = 8.85419 \times 10^{-12} \text{ C}^2 / \text{N m}^2$ is the permittivity of free space. Substituting these values into Eq. (9.4), we get

$$c = 2.99792 \times 10^8 \text{ m/s}$$

which is the same as the speed of light in vacuum. This strongly supports the conclusion that light is an electromagnetic wave. The speed of em wave or light in any other medium of permittivity ϵ and permeability μ is

$$v = \frac{1}{\sqrt{\epsilon \mu}} = \frac{1}{\sqrt{K \epsilon_0 \mu_r \mu_0}} = \frac{c}{\sqrt{K \mu_r}} \quad (9.5)$$

where K is the dielectric constant (or relative permittivity) of the medium and μ_r is the relative permeability.

Electric and magnetic fields in an em wave in free space are always perpendicular to each other and also perpendicular to the direction of travel of the wave. Thus, electromagnetic waves are transverse waves. As mentioned earlier, the ratio of the magnitudes of the electric to the magnetic field in an electromagnetic wave equals to the speed of light.

Electromagnetic waves carry energy as they travel through space and this energy is shared equally by the electric and magnetic fields. The technological importance of em waves lies in its capability to carry energy from one place to another. They not only transmit energy from radio and TV stations to our homes, but also carry energy from the sun to the earth.

Electromagnetic waves transport linear momentum as well. When electromagnetic waves strike a surface, a pressure is exerted on the

surface. If the total energy transferred to a surface in time t is U , it can be shown that the magnitude of the total momentum delivered to this surface is,

$$p = \frac{U}{c} \quad (\text{complete absorption}) \quad (9.6)$$

When the sun shines on your hand, you feel the energy being absorbed from the electromagnetic waves (your hands get warm). Electromagnetic waves also transfer momentum to your hand but because c is very large, the amount of momentum transferred is extremely small and you do not feel the pressure. In 1903, the American scientists Nicols and Hull succeeded in measuring radiation pressure of visible light and verified Eq. (9.6). It was found to be of the order of $7 \times 10^{-6} \text{ N/m}^2$.

Thus, on a surface of area 10 cm^2 , the force due to radiation is only about $7 \times 10^{-9} \text{ N}$.

Example 9.1 A plane electromagnetic wave of frequency 25 MHz travels in free space along the x -direction. At a particular point in space and time, $\mathbf{E} = 6.3 \hat{j} \text{ V/m}$. What is \mathbf{B} at this point?

Answer The magnitude of \mathbf{B} and \mathbf{E} are related by

$$B = \frac{E}{c} = \frac{6.3 \text{ V/m}}{3 \times 10^8 \text{ m/s}} = 2.1 \times 10^{-8} \text{ T}$$

To find the direction, we note that \mathbf{E} is along y -direction and the wave propagate along x -axis. Therefore, \mathbf{B} should be in a direction perpendicular to both x - and y -axes. Using vector algebra, $\mathbf{E} \times \mathbf{B}$ should be along x -direction. Since,

$$(+\hat{j}) \times (+\hat{k}) = \hat{i}, \mathbf{B} \text{ is along the } z\text{-direction.}$$

$$\text{Thus, } \mathbf{B} = 2.1 \times 10^{-8} \hat{k} \text{ T}$$

Example 9.2 The magnetic field in a plane electromagnetic wave is given by $B_y = 2 \times 10^{-7} \sin(0.5 \times 10^3 x + 1.5 \times 10^{11} t) \text{ T}$. (a) What is the wavelength and frequency of the wave? (b) Write an expression for the electric field.

Answer

(a) Comparing the given equation with

$$B_y = B_0 \sin \left[2\pi \left(\frac{x}{\lambda} + \frac{t}{T} \right) \right]$$

$$\text{We get, } \lambda = \frac{2\pi}{0.5 \times 10^3} \text{ m} = 1.26 \text{ cm,}$$

$$\frac{1}{T} = \nu = (1.5 \times 10^{11}) / 2\pi = 23.9 \text{ GHz.}$$

(b) $B_0 = 2 \times 10^{-7} \text{ T} \therefore E_0 = B_0 c = 6 \times 10^1 \text{ V/m}$
The electric field component is perpendicular to the direction of propagation and the direction of magnetic field. Therefore,
 $E_z = 60 \sin(0.5 \times 10^3 x + 1.5 \times 10^{11} t) \text{ V/m.}$ ◀

Example 9.3 Light with an energy flux of 18 watts/cm^2 falls on a non-reflecting surface at normal incidence. If the surface has an area of 20 cm^2 , find the average force exerted on the surface during a 30 minute time span.

Answer The total energy falling on the surface is

$$U = (18 \text{ W/cm}^2) (30 \times 60) (20 \text{ cm}^2) \\ = 6.48 \times 10^5 \text{ J}$$

Therefore, the total momentum delivered is

$$p = \frac{U}{c} = \frac{6.48 \times 10^5 \text{ J}}{3 \times 10^8 \text{ m/s}} \\ = 2.16 \times 10^{-3} \text{ kg m/s}$$

The average force exerted on the surface is

$$F = \frac{p}{t} = \frac{2.16 \times 10^{-3}}{0.18 \times 10^4} = 1.2 \times 10^{-6} \text{ N.}$$

How will your result be modified if the surface is a perfect reflector? In this case, the change of momentum will be twice the above value so the force will be twice ($2.4 \times 10^{-6} \text{ N}$). ◀

Example 9.4 Obtain an expression for (a) energy density and (b) intensity of radiation of an electromagnetic wave.

Answer In previous chapters on static electricity and magnetism, we discussed the idea of electrostatic and magnetic energy density. We also noted that energy can be stored in space wherever electric and magnetic fields are present. In empty space, the energy density u consists of electric and magnetic contributions :

$$u = u_E + u_M = \frac{1}{2} \epsilon_0 E^2 + \frac{B^2}{2\mu_0}$$

where E and B are the static electric and magnetic fields. That is, the value of E and B are constant and do not vary with time. For an

electromagnetic wave, E and B are sinusoidally varying functions of space and time. The above expressions are still valid for em waves if E and B are replaced by their rms values.

For a plane wave of electromagnetic radiation, B and E are related by,

$$B_{\text{rms}} = \frac{E_{\text{rms}}}{c}$$

Therefore,

$$\begin{aligned} u &= \frac{1}{2} \epsilon_0 E_{\text{rms}}^2 + \frac{1}{2 \mu_0} \frac{E_{\text{rms}}^2}{c^2} \\ &= \frac{1}{2} \epsilon_0 E_{\text{rms}}^2 + \frac{\epsilon_0 \mu_0}{2 \mu_0} E_{\text{rms}}^2 \end{aligned}$$

since $c^2 = \frac{1}{\epsilon_0 \mu_0}$

or $u = \frac{1}{2} \epsilon_0 E_{\text{rms}}^2 + \frac{1}{2} \epsilon_0 E_{\text{rms}}^2 = \epsilon_0 E_{\text{rms}}^2 \quad (9.7)$

(b) The intensity of radiation is defined as the amount of energy passing through unit area in unit time:

$$I = \frac{\text{Energy / time}}{\text{Area}} = \frac{\text{Power}}{\text{Area}}$$

u is the energy density, i.e., energy in unit volume of space. In one second, em waves travel a distance of c . Energy passing through a unit area per unit time $= u c$. Therefore,

$$I = u c = \epsilon_0 c E_{\text{rms}}^2 \quad (9.8)$$

Example 9.5 Calculate the electric and magnetic fields produced by the radiation coming from a 100 watt bulb at a distance of 3 m. Assume that the efficiency of the bulb is 2.5% and it is a point source.

Answer The bulb, as a point source, radiates light in all directions uniformly. At a distance of 3 m, the surface area of the surrounding sphere is

$$A = 4\pi r^2 = 4\pi(3)^2 = 113\text{ m}^2$$

The intensity at this distance is

$$\begin{aligned} I &= \frac{\text{Power}}{\text{Area}} = \frac{100\text{ W} \times 2.5\%}{113\text{ m}^2} \\ &= 0.022\text{ W/m}^2 \end{aligned}$$

Half of this intensity is provided by the electric field and half by the magnetic field.

$$\frac{1}{2} I = \frac{1}{2} (\epsilon_0 E_{\text{rms}}^2 c) = \frac{1}{2} (0.022\text{ W/m}^2)$$

$$\begin{aligned} E_{\text{rms}} &= \sqrt{\frac{0.022}{(8.85 \times 10^{-12})(3 \times 10^8)}}\text{ V/m} \\ &= 2.9\text{ V/m} \end{aligned}$$

The value of E found above is the root mean square value of the electric field. Since the electric field in a light beam is sinusoidal, the peak electric field, E_0 is

$$\begin{aligned} E_0 &= \sqrt{2} E_{\text{rms}} = \sqrt{2} \times 2.9\text{ V/m} \\ &= 4.07\text{ V/m} \end{aligned}$$

Thus, you see that the electric field strength of the light that you use for reading is fairly large. Compare it with electric field strength of TV or FM waves which is of the order of a few microvolts per metre.

Now, let us calculate the strength of the magnetic field. It is

$$\begin{aligned} B_{\text{rms}} &= \frac{E_{\text{rms}}}{c} = \frac{2.9\text{ V m}^{-1}}{3 \times 10^8\text{ m s}^{-1}} \\ &= 9.6 \times 10^{-9}\text{ T} \end{aligned}$$

Again, since the field in the light beam is sinusoidal, the peak magnetic field is $\sqrt{2} B = 1.4 \times 10^{-8}\text{ T}$. Note that although the power in the magnetic field is equal to the power in the electric field, the magnetic field strength is evidently very weak.

9.3 ELECTROMAGNETIC SPECTRUM

At the time Maxwell predicted the existence of electromagnetic waves, the only familiar em waves were visible light waves. The existence of ultraviolet and infrared waves was barely established. By the end of the nineteenth century, X-rays and gamma rays had also been discovered. We now know that, em waves include visible light waves, X-rays, gamma rays, radio waves, microwaves, ultraviolet and infrared waves. The classification of em waves according to frequency is the electromagnetic spectrum (Fig. 9.7). There is no sharp division between one kind of wave and the next. The classification is based roughly on how the waves are produced and/or detected.

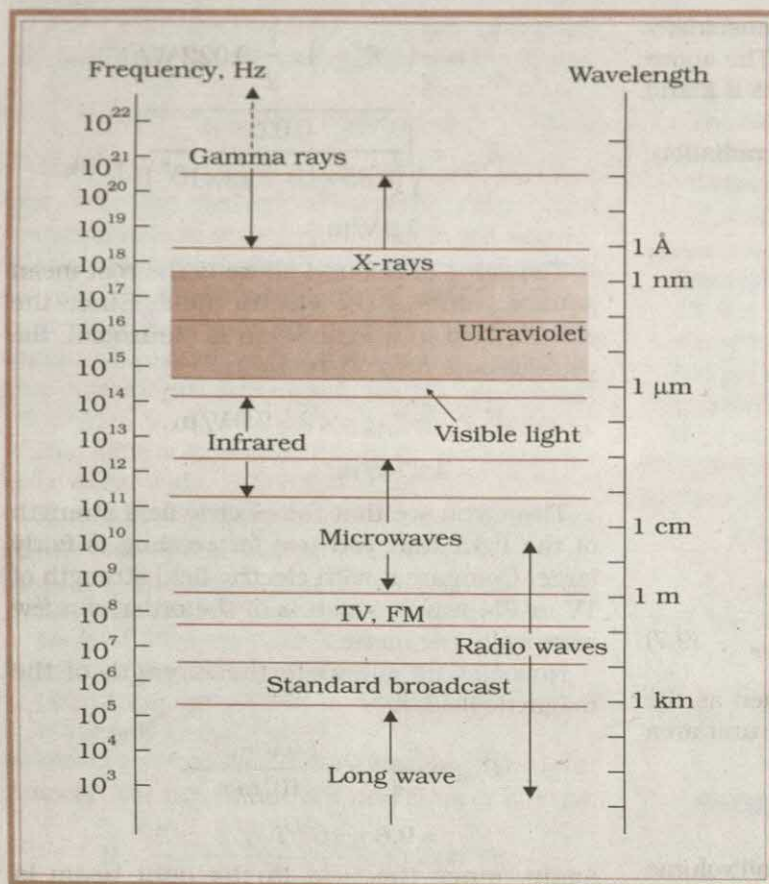


Fig. 9.7 The electromagnetic spectrum. Note the overlap between one type of wave and the next. There is no sharp division between the different regions of em spectrum.

We briefly describe these different types of em waves, in order of decreasing wavelength.

Radio waves

Radio waves are produced by the accelerated motion of charges in conducting wires. They are used in radio and television communication systems.

They are generally in the frequency range from 500 kHz to about 1000 MHz. The AM (amplitude modulated) band is from 530 kHz to 1710 kHz. Higher frequencies upto 54 MHz are used for 'short wave' bands. TV waves range from 54 MHz to 890 MHz. The FM (frequency modulated) radio band extends from 88 MHz to 108 MHz. Cellular phones use radio waves to transmit voice communication in the ultrahigh frequency (UHF) band. How these waves are transmitted and received is described in Chapter 16.

Microwaves

Microwaves (short-wavelength radio waves), with frequencies in the gigahertz (GHz) range, are produced by special vacuum tubes (called klystrons, magnetrons and Gunn diodes). Due to their short wavelengths, they are suitable for the radar systems used in aircraft navigation. Radar also provides the basis for the speed guns used to time fast balls, tennis serves, and automobiles. Microwave ovens are an interesting domestic application of these waves.

Infrared waves

Infrared waves are produced by hot bodies and molecules. This band lies adjacent to the low-frequency or long-wave length end of the visible spectrum. Infrared waves are sometimes referred to as *heat waves*. This is because water molecules present in most materials readily absorb infrared waves (Many other molecules, e.g., CO_2 , NH_3 , also absorb infrared waves). After absorption, their thermal motion increases, that is, they heat up and heat their surroundings. Infrared

lamps are used in physical therapy. Infrared radiation also plays an important role in maintaining the earth's warmth or average temperature through the greenhouse effect. Incoming visible light (which passes relatively easily through the atmosphere) is absorbed by the earth's surface and re-radiated as infrared (longer wavelength) radiations. This radiation is trapped by greenhouse gases such as carbon dioxide and water vapour.

Visible light

It is the most familiar form of electromagnetic waves. It is the part of the spectrum that is detected by the human eye. It runs from about 4×10^{14} Hz to about 7×10^{14} Hz or a wavelength range of about 700-400 nm. Visible light emitted or reflected from objects around us provides us information about the world.

Our eyes are sensitive to this range of wavelengths. Different animals are sensitive to different range of wavelengths. For example, snakes can detect infrared waves, and the 'visible' range of many insects extends well into the ultraviolet.

Ultraviolet light

It covers wavelengths ranging from about 4×10^{-7} m (400 nm) down to 6×10^{-10} m (0.6 nm). Ultraviolet (UV) radiation is produced by special lamps and very hot bodies. The sun is an important source of ultraviolet light. But fortunately, most of it is absorbed in the ozone layer in the atmosphere at an altitude of about 40-50 km. UV light in large quantities has harmful effects on humans. Since ozone layer plays a protective role, its depletion by chlorofluorocarbons (CFCs) gas (such as freon) is a matter of international concern.

X-rays

Beyond the UV region of the electromagnetic spectrum lies the X-ray region. We are familiar with X-rays because of its medical applications. It covers wavelengths from about 10^{-8} m (10 nm) down to 10^{-13} m (10^{-4} nm). One common way to generate X-rays is to bombard a metal target by high energy electrons. X-rays are used as a

diagnostic tool in medicine and as a treatment for certain forms of cancer. Because X-rays damage or destroy living tissues and organisms, care must be taken to avoid unnecessary or over exposure. X-rays are discussed in Chapter 13.

Gamma rays

They lie in the upper frequency range of the electromagnetic spectrum and have wavelengths ranging from about 10^{-10} m to less than 10^{-14} m. This high frequency radiation is produced in nuclear reactions and also emitted by radioactive nuclei. You will encounter them in Chapter 14.

9.4 PROPAGATION OF ELECTROMAGNETIC WAVES IN ATMOSPHERE

Before we discuss the propagation of em waves in the atmosphere, it is necessary to know a few things about the atmosphere and its various layers. The atmosphere is the gaseous envelope surrounding our earth. It is retained to the earth due to gravitational attraction. As we go up, the air thins out gradually and air pressure decreases. The atmosphere can be divided into various layers as shown in Fig. 9.8. The layers are known by different names and with tops denoted by pauses.

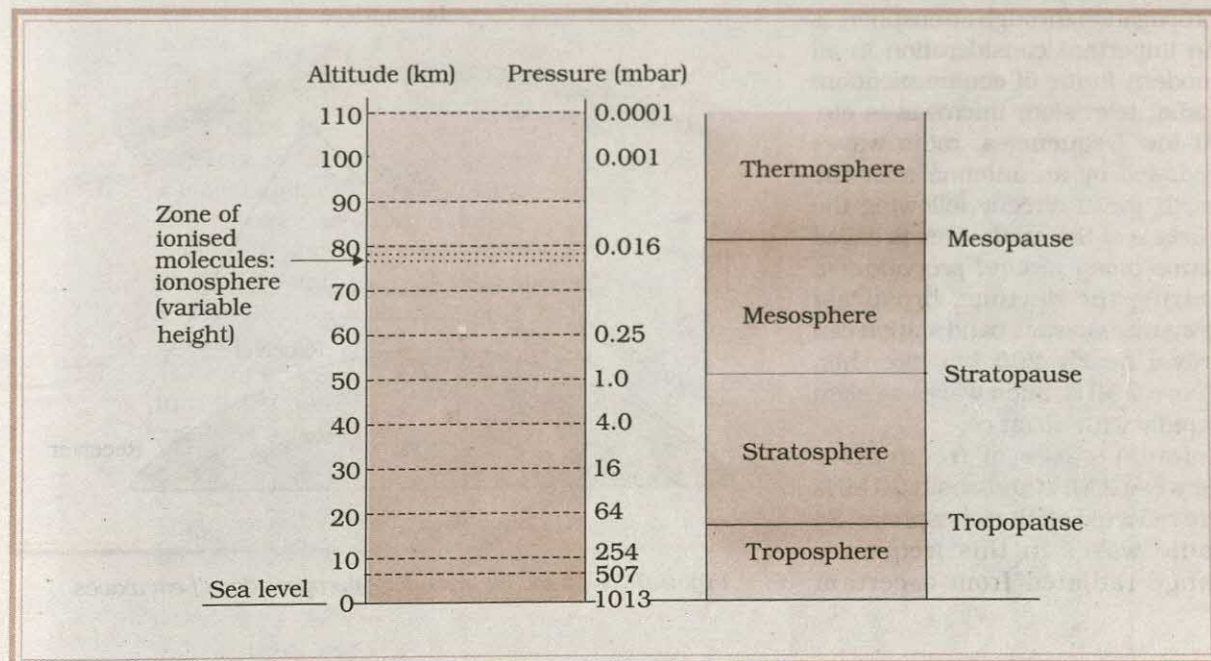


Fig. 9.8 The earth's atmosphere.

The *troposphere* includes the layer close to the earth and extends upto about 12 km. This layer is responsible for all the important weather phenomena affecting our environment. The next layer, called the *stratosphere*, extends from about 10-16 km to about 50 km. The *mesosphere* extends from about 50 km to about 80 km. The *thermosphere* extends from 80 km to the edge of the atmosphere. It receives energy directly from the solar radiation. The ozone layer is in the lower stratosphere and extends from 15 km to about 30 km. This ozone results from the dissociation of molecular oxygen by solar ultraviolet radiation in the upper atmosphere. Except for the layer in the upper atmosphere, called *ionosphere*, which is composed partly of electrons and positive ions, the rest of the atmosphere is composed mostly of neutral molecules.

The atmosphere is transparent to visible radiation and we can see the sun and the stars through it clearly. However, most infrared radiation is not able to pass through, as it is absorbed by the atmosphere. Low lying clouds in the atmosphere also prevent infrared radiation from passing through. The ozone layer blocks the passage of ultraviolet radiation from the sun.

The behaviour of em waves of wavelength 10^{-3} m and higher (called radio waves) in their propagation through atmosphere is an important consideration in all modern forms of communication: radio, television, microwaves etc. At low frequencies, radio waves radiated by an antenna near the earth travel directly following the surface of the earth. This is called *wave along ground propagation*. During the daytime, broadcast from medium waveband station can travel nearly 200 km like this. Above 2 MHz, such waves weaken rapidly with distance.

Radio waves of frequencies between 2 MHz and about 20 MHz are reflected off the ionosphere. So radio waves in this frequency range radiated from a certain

point and after being reflected by the ionosphere, can be received at another point on the surface. This is known as *sky wave* or ionospheric propagation. In this way, radio waves can travel very large distances and can even travel round the earth.

Ionosphere does not help in the propagation of waves of frequencies higher than 30 MHz. Television signals have frequencies in the 100 - 200 MHz range and penetrate ionosphere (no reflection) and therefore their propagation is not possible through the sky wave. Reception of such waves is possible only if the receiver antenna directly intercepts the signals. Thus, television broadcasts are made from tall antenna to get larger coverage. This is *space wave* propagation. Radiowaves with frequencies higher than television signals are the microwaves. In recent times, microwaves have revolutionised telecommunications. The signals (in this range) from the broadcasting station are beamed towards a geostationary satellite, which in turn broadcasts it back to the earth. In this way, signals can be propagated over a large part of the earth's surface.

You will learn more about radio and TV communication in Chapter 16.

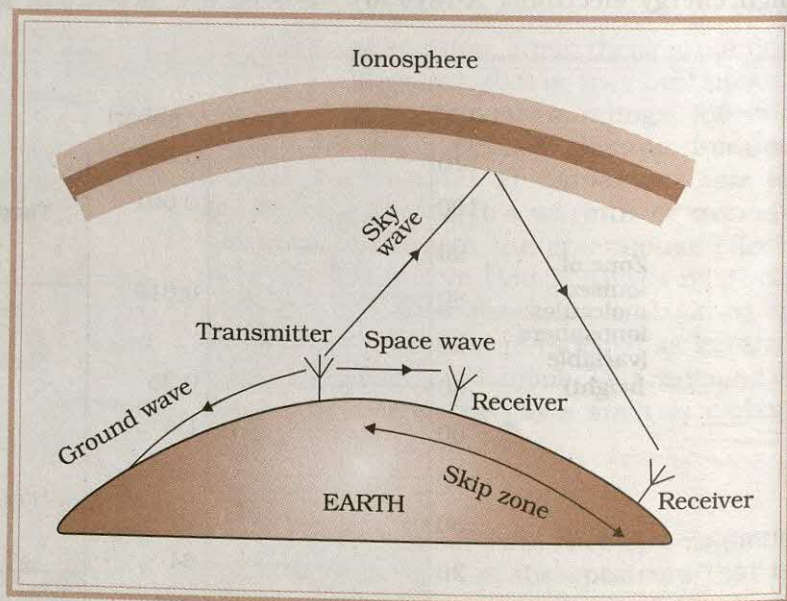


Fig. 9.9 The three main modes of propagation of em waves.

SUMMARY

1. Maxwell proved that a magnetic field is produced not only by a current, but also by a time-varying electric field. Maxwell's equations represent all the basic laws of electromagnetism.
2. An accelerating charge produces electromagnetic waves. An electric charge oscillating harmonically with frequency ν , produces electromagnetic waves of the same frequency ν . An electric dipole is a basic source of electromagnetic waves.
3. Electromagnetic waves with wavelength of the order of a few meters were first produced and detected in the laboratory by Hertz in 1887. He thus verified a basic prediction of Maxwell's equations.
4. Electric and magnetic fields oscillate sinusoidally in space and time in an electromagnetic wave, giving rise to and sustaining each other. The oscillating electric and magnetic fields, \mathbf{E} , and \mathbf{B} are perpendicular to each other, and to the direction of propagation of the electromagnetic wave. For a wave of frequency ν , wavelength λ , propagating along x -direction, we have

$$\begin{aligned}
 E &= E_y(t) = E_0 \sin(kx - \omega t) \\
 &= E_0 \sin\left[2\pi\left(\frac{x}{\lambda} - \nu t\right)\right] = E_0 \sin\left[2\pi\left(\frac{x}{\lambda} - \frac{t}{T}\right)\right] \\
 B &= B_z(t) = B_0 \sin(kx - \omega t) \\
 &= B_0 \sin\left[2\pi\left(\frac{x}{\lambda} - \nu t\right)\right] = B_0 \sin\left[2\pi\left(\frac{x}{\lambda} - \frac{t}{T}\right)\right]
 \end{aligned}$$

They are related by $E_0/B_0 = c$.

5. The speed c of electromagnetic wave in vacuum is related to μ_0 and ϵ_0 (the free space permeability and permittivity constants) as follows: $c = 1 / \sqrt{\mu_0 \epsilon_0}$. The value of c equals the speed of light obtained from optical measurements. Light is an electromagnetic wave; c is, therefore, also the speed of light. Electromagnetic waves other than light also have the same velocity c in free space.

The speed of light, or of electromagnetic waves in a material medium is given by $v = 1 / \sqrt{\mu \epsilon}$

where μ is the permeability of the medium and ϵ its permittivity..

6. Electromagnetic waves carry energy as they travel through space and this energy is shared equally by the electric and magnetic fields. Electromagnetic waves transport momentum as well. When these waves strike a surface, a pressure is exerted on the surface. If total energy transferred to a surface in time t is U , total momentum delivered to this surface is $p = U/c$.
7. The spectrum of electromagnetic waves stretches, in principle, over an infinite range of wavelengths. Different regions are known by different names; γ -rays, X-rays, ultraviolet rays, visible light, infrared rays, microwaves and radio waves in order of increasing wavelength from 10^{-2} Å or 10^{-12} m to 10^6 m.

They interact with matter via their electric and magnetic fields which set in oscillation charges present in all matter. The detailed interaction and so the mechanism of absorption, scattering etc., depend on the wavelength of the em wave, and the nature of the atoms and molecules in the medium.

8. A medium of particular importance is the atmosphere. It is transparent to visible light. The ozone layer absorbs harmful ultraviolet light. Infrared waves both from the sun and re-emitted from the earth are partially absorbed by the atmospheric gases such as carbon dioxide, and by clouds. The energy balance near the earth's surface is affected by man's activities, and can lead to changes such as global warming or ozone depletion. The propagation of radio waves depends on the wavelength of the waves. Medium frequency (MF) waves (300 kHz-3 MHz) are largely absorbed and the high frequency (HF) waves (3-30 MHz) are reflected back by the ionosphere. In the range 30 MHz to 3 GHz, waves are transmitted from one place to another either by direct line-of-sight using tall towers, or by beaming to artificial satellites and rebroadcasting from there.

POINTS TO PONDER

1. The basic difference between various types of electromagnetic waves lies in their wavelengths or frequencies since all of them travel through vacuum with the same speed. Consequently, the waves differ considerably in their mode of interaction with matter.
2. Accelerated charged particles radiate electromagnetic waves. The wavelength of the electromagnetic wave is often correlated with the characteristic size of the system that radiates. Thus, gamma radiation, having wavelength of 10^{-14} m to 10^{-15} m, typically originate from an atomic nucleus. X-rays are emitted from heavy atoms. Radio waves are produced by accelerating electrons in a circuit. A transmitting antenna can most efficiently radiate waves having a wavelength of about the same size as the antenna. Visible radiation emitted by atoms is, however, much longer in wavelength than atomic size.
3. The oscillating fields of an electromagnetic wave can accelerate charges and can produce oscillating currents. Therefore, an apparatus designed to detect electromagnetic waves is based on this fact. Hertz original 'receiver' worked in exactly this way. The same basic principle is utilised in practically all modern receiving devices. High frequency electromagnetic waves are detected by other means based on the physical effects they produce on interacting with matter.
4. Infrared waves, with frequencies lower than those of visible light, vibrate not only the electrons, but entire atoms or molecules of a substance. This vibration increases the internal energy and consequently, the temperature of the substance. This is why infrared waves are often called *heat waves*.
5. The centre of sensitivity of our eyes coincides with the centre of the wavelength distribution of the sun. It is because humans have evolved with visions most sensitive to the strongest wavelengths from the sun.
6. The great technological import of electromagnetic waves stems from their capability to transfer energy from one place to another. The radio and TV signals from broadcasting stations carry energy. Light carries energy from the sun to the earth, thus making life possible on the earth.

EXERCISES

- 9.1 What physical quantity is the same for X-rays of wavelength 10^{-10} m, red light of wavelength 6800 Å and radiowaves of wavelength 500m?
- 9.2 A plane electromagnetic wave travels in vacuum along z-direction. What can you say about the directions of its electric and magnetic field vectors? If the frequency of the wave is 30 MHz, what is its wavelength?
- 9.3 A radio can tune in to any station in the 7.5 MHz to 12 MHz band. What is the corresponding wavelength band?
- 9.4 A charged particle oscillates about its mean equilibrium position with a frequency of 10^9 Hz. What is the frequency of the electromagnetic waves produced by the oscillator?
- 9.5 The amplitude of the magnetic field part of a harmonic electromagnetic wave in vacuum is $B_0 = 510$ nT. What is the amplitude of the electric field part of the wave?
- 9.6 Suppose that the electric field amplitude of an electromagnetic wave is $E_0 = 120$ N/C and that its frequency is $\nu = 50.0$ MHz. (a) Determine, B_0 , ω , k and λ . (b) Find expressions for \mathbf{E} and \mathbf{B} .
- 9.7 The terminology of different parts of the electromagnetic spectrum is given in the text. Use the formula $E = h\nu$ (for energy of a quantum of radiation: photon) and obtain the photon energy in units of eV for different parts of the em spectrum. In what way are the different scales of photon energies that you obtain related to the sources of electromagnetic radiation?
- 9.8 In a plane em wave, the electric field oscillates sinusoidally at a frequency of 2.0×10^{10} Hz and amplitude 48 V m $^{-1}$.
 (a) What is the wavelength of a wave?
 (b) What is the amplitude of the oscillating magnetic field?
 (c) Show that the average energy density of the \mathbf{E} field equals the average energy density of the \mathbf{B} field. [$c = 3 \times 10^8$ m s $^{-1}$].

ADDITIONAL EXERCISES

- 9.9 Suppose that the electric field part of an electromagnetic wave in vacuum is $\mathbf{E} = [(3.1 \text{ N/C}) \cos [(1.8 \text{ rad/m}) y + (5.4 \times 10^6 \text{ rad/s}) t]] \hat{i}$
 (a) What is the direction of propagation?
 (b) What is the wavelength λ ?
 (c) What is the frequency ν ?
 (d) What is the amplitude of the magnetic field part of the wave?
 (e) Write an expression for the magnetic field part of the wave.
- 9.10 About 5% of the power of a 100 W light bulb is converted to visible radiation. What is the average intensity of visible radiation
 (a) at a distance of 1m from the bulb?
 (b) At a distance of 10 m?
 Assume that the radiation is emitted isotropically and neglect reflection.
- 9.11 Use the formula $\lambda_m T = 0.29$ cm K to obtain the characteristic temperature ranges for different parts of the em spectrum. What do the numbers that you obtain tell you?
- 9.12 Magnetic field lines can never emanate from a point nor end on a point. Yet the field lines outside a bar magnet do seem to start from the North pole and end on the South pole. Does the second fact contradict the first? Explain.

9.13 Given below are some famous numbers associated with electromagnetic radiation in different contexts in physics. State the part of the em spectrum to which each belongs.

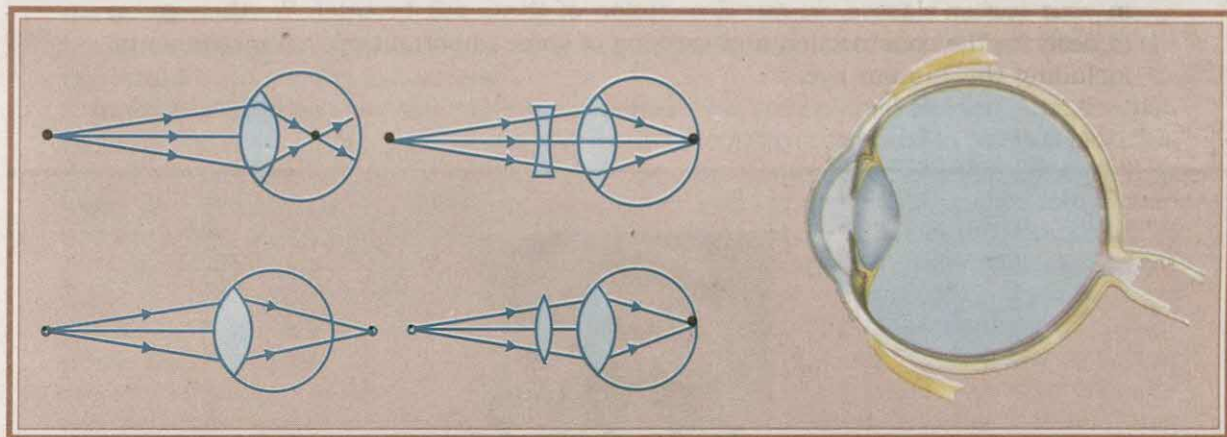
- (a) 21 cm (wavelength emitted by atomic hydrogen in interstellar space).
- (b) 1057 MHz (frequency of radiation arising from two close energy levels in hydrogen; known as Lamb shift).
- (c) 2.7 K [temperature associated with the isotropic radiation filling all space-thought to be a relic of the 'big-bang' origin of the universe].
- (d) 5890 Å - 5896 Å [double lines of sodium]
- (e) 14.4 keV [energy of a particular transition in ^{57}Fe nucleus associated with a famous high resolution spectroscopic method (Mössbauer spectroscopy)].

9.14 Answer the following questions:

- (a) Long distance radio broadcasts use short-wave bands. Why?
- (b) It is necessary to use satellites for long distance TV transmission. Why?
- (c) Optical and radiotelescopes are built on the ground but X-ray astronomy is possible only from satellites orbiting the earth. Why?
- (d) The small ozone layer on top of the stratosphere is crucial for human survival. Why?
- (e) If the earth did not have an atmosphere, would its average surface temperature be higher or lower than what it is now?
- (f) Some scientists have predicted that a global nuclear war on the earth would be followed by a severe 'nuclear winter' with a devastating effect on life on earth. What might be the basis of this prediction?

CHAPTER TEN

RAY OPTICS AND OPTICAL INSTRUMENTS



10.1 INTRODUCTION

We all know that nature has endowed the human eye (retina) with the sensitivity to detect a small 'window' of the electromagnetic spectrum. Electromagnetic radiation belonging to this region of the spectrum (wavelength of about 400 nm to 700 nm) is called light. It is mainly through light and the sense of vision that we know and interpret the world around us.

There are two things that we can intuitively mention about light from common experience. First, that it travels with enormous speed and second, that it travels in a straight line. It took some time for people to realise that the speed of light is finite and measurable. We shall not discuss the various optical methods by which the speed of light was determined. Its presently accepted value is: $c = 2.99792458 \times 10^8 \text{ m s}^{-1}$. For many purposes, it suffices to take $c = 3 \times 10^8 \text{ m s}^{-1}$. In fact the precision of this measurement is now so great that the standard metre is now **defined** to be the length of the path travelled by light in vacuum during a time interval of $1/299792458$ of a second. The speed of light in vacuum is the highest speed attainable in nature. No physical signal or message can travel with a speed greater than c . Further, the speed of light in vacuum is independent of the relative motion between the source and the observer. This fact underlines Einstein's special theory of relativity. You already know that the speed of light in any material medium is c/n , where n is the refractive index of the medium with respect to vacuum. Since $n > 1$, the speed in a medium is always less than c .

The intuitive notion that light travels in a straight line seems to contradict what we have learnt in the previous Chapter, namely, that light is an electromagnetic wave of wavelength belonging to the visible part of the spectrum. How to reconcile the two facts? The answer is that the wavelength of light is very small compared to the size of ordinary objects that we encounter commonly (generally of the order of a few cm or larger). In this situation, as you will learn in Chapter 11, a light wave can be considered to travel from one point to another along a straight line path joining them. This path is called a ray of light; and a bundle of such rays constitutes a beam of light.

In this Chapter, we consider the phenomena of reflection, refraction and dispersion of light, using the ray picture of light. Using the basic laws of reflection and refraction, we shall study the image formation by plane and spherical reflecting and refracting surfaces. You have learnt some aspects of these topics in your earlier classes, so our description of these will be brief. We then go on to describe the construction and working of some important optical instruments, including the human eye.



Albert Abraham Michelson **(1852-1931)**

An outstanding American experimenter with a long and distinguished career. He measured the speed of light with steadily improving accuracy, ultimately reaching within 1 kilometer per second of the presently accepted value. His experiment with Morley failed to find changes in the speed of light produced by motion with respect to the ether and thus laid the foundation for special relativity. The interference arrangement which he invented for this purpose was put to good use in spectroscopy, giving a wavelength resolution of a thousandth of a nanometre. His work with Pease on the coherence of starlight collected from two mirrors upto 20 ft. apart gave the first measurement of the angle subtended by any star (less than 10^{-6} radian!).

NEWTON'S PARTICLE MODEL OF LIGHT

Newton's fundamental contributions to mathematics, mechanics, and gravitation often blind us to his deep experimental and theoretical study of light. He proposed that light energy is concentrated in tiny particles called 'corpuscles'. With his understanding of mechanics, he could come up with a simple model of reflection and refraction. It is a common observation that a ball bouncing from a smooth plane surface obeys the laws of reflection. When this is an elastic collision, the magnitude of the velocity remains the same. Because the surface is smooth, there is no force acting parallel to the surface, so the component of momentum in this direction also remains the same. Only the component perpendicular to the surface, i.e., the normal component of the momentum, gets reversed in reflection.

Coming to refraction, Newton postulated that the speed of the corpuscles was greater in water or glass than in air. On entering the medium, the momentum component parallel to the surface would remain the same (as in the case of reflection). The normal component would become greater because of a force acting perpendicular to the surface. Hence, the ray would bend toward the normal.

In the field of optics, Newton – the experimenter, was greater than Newton – the theorist. He himself observed many phenomena which are difficult to understand in terms of particles. We will see one example in the next Chapter, viz., the colours shown by thin films of oil on water. But there is a much simpler example, which every child who has looked into a pond has seen. This is the property of partial reflection. One sees one's face in the pond, but also the bottom of the pond. This meant, to Newton, that some of the corpuscles which fall on the water get reflected, and some get transmitted. But what property could distinguish these two kinds of corpuscles? Newton had to postulate some kind of unpredictable, chance phenomenon which decided whether an individual corpuscle would be reflected or not. In other experiments, however, the corpuscles behave as if they are identical. This basic difficulty occurs because particles are indivisible. It does not occur in a pure wave picture. An incoming wave can divide into two weaker waves at the boundary between air and water.

You will see in Chapter 12 that Newton's difficulty still remains unresolved. Not just light, but also electrons, show particle-like as well as wave-like behaviour. The fate of individual electrons is governed by probability.

LIGHT SOURCES AND PHOTOMETRY

As you saw in Chapter 13 of Class XI textbook, any hot body will emit electromagnetic waves. This radiation occupies a wide range of wavelengths, but the energy is greatest near a value which depends inversely on the temperature. A good example is the Sun, whose surface is at a temperature of about 5500 K. The graph of energy emitted as a function of wavelength has a peak at $\lambda = 550$ nm, which is green, and in the middle of the visible region of the spectrum. Hence, most of the energy from the Sun can be perceived by the human eye. (This is, of course, not an accident or coincidence — human and other eyes evolved on earth, surrounded by sunlight!).

We call a source like the sun thermal. The hot filament of an incandescent lamp also emits a broad spectrum with a peak, like the Sun, and it is also a thermal source. However, its temperature is less than 3000 K, the melting point of tungsten. Therefore, the peak is at a wavelength longer than 1100 nm, i.e., in the near infrared part of the electromagnetic spectrum. Only a small fraction of the total energy is at wavelengths shorter than 700 nm i.e., in the visible range. Most of this is in the red and yellow, and only very little is in the blue region.

The other familiar kind of source is the 'tube light', i.e., the mercury vapour lamp. In this case, the energy supplied is not in the form of heat. The external voltage across the

tube creates an electric field from which free electrons in the tube pick up kinetic energy. These collide with mercury atoms which go to excited states (Chapter 13). When they come back to the state of lowest energy, radiation of specific wavelengths is emitted, not a continuous range. This kind of source is called a fluorescent lamp for a specific reason. The wall of a tube is coated with a material which absorbs ultraviolet radiation from the mercury atoms and reemits visible radiation — a process called fluorescence. The sodium vapour lamp, often used for street lighting, emits a pair of wavelengths in the orange region of the spectrum. No one should buy clothes in a shop lit with sodium lamps, and even mercury lamps, because their strong violet and blue lines give a different appearance to coloured clothes, when compared to daylight.

The word 'photometry' means the measurement of light. From the physicist's viewpoint, the strength of a source of electromagnetic radiation is measured by the amount of energy that it emits in a unit time. The appropriate unit would be J/s or W. This is called the luminosity and we denote it by L . If we now have a detector of radiation of unit area, placed at a distance r from this source, how much radiation does it receive per second? Constructing a sphere of radius r around the source, it has an area $4\pi r^2$. For an 'isotropic' source — one emitting equally in all directions — the entire luminosity of L J/s has to pass through this sphere. Hence, the energy per unit area per unit time — called 'flux' F , is given by

$$F = \frac{L}{4\pi r^2}$$

Note that the flux of sunlight at the Earth is 1.4×10^3 W/m². The luminosity, L , of sun turns out to be 4×10^{26} W. A large power station produces about 10^9 W.

In astronomy and other subjects, it may be necessary to deal with the luminosity and flux per unit interval of frequency or wavelength, as a function of wavelength. This naturally gives us more information. However, the geometrical argument given above is still true for these quantities.

The expression for flux, F , given above is called the 'inverse square law' or 'Lambert's Law'. We see its operation in everyday life. A weak bulb on a table lamp one metre above your notebook will light it up in the same way as a nine times stronger bulb placed 3 m away from the book.

In situations such as room or street lighting, we are not concerned just with the energy of light, but how effective it is in creating a sensation in the human eye. The same power, measured in joules per second (watts), will be perceived by the eye as having different 'brightness' at different wavelengths. The strength of a light source is hence measured in a new unit called the *lumen*. For a given spectrum of the source, one can relate lumens to watts. This conversion factor is at its maximum when the wavelength is that at which the human eye is most sensitive, about 555 nm. One watt of power at this wavelength corresponds to 683 lumens. For comparison, an incandescent (filament) bulb of 40 W gives only about 450 lumens. This illustrates the point made earlier that the human eye is sensitive only to a small fraction of the radiation emitted by such a lamp, because most of the energy is in infrared radiation. The tube light is much better, giving nearly 2000 lumens for 40 W. A unit which is related to the lumen is the candela. The candela is defined as one lumen per steradian. It is a unit of 'luminous intensity', i.e., luminous power (in lumens) per unit solid angle.

'Steradian' is a unit of solid angle analogous to the radian for plane angles. Just as an arc equal to the radius subtends one radian, an area on a sphere of radius r , (equal to r^2), subtends a solid angle of one steradian (sr), and the full sphere as 4π steradians. You have encountered this earlier in the discussion of Gauss's theorem in electrostatics (Chapter 1).

When one has to read a book, or drive on a road, the relevant quantity is the luminous power Φ reaching unit area of the surface. This is called 'illuminance' and is measured in lumens per square metre (also called lux). A source of Φ lumens will produce an illuminance of $\Phi/4\pi r^2$ lux on a surface placed at a distance r , held perpendicular to the rays.

10.2 REFLECTION OF LIGHT BY SPHERICAL MIRRORS

The laws of reflection by a plane mirror are familiar: the angle of incidence (angle between incident ray and the normal to the mirror) equals the angle of reflection (angle between reflected ray and the normal). These laws are applied at every point on the surface of a spherical mirror. The normal in this case is to be taken as normal to the tangent to the surface at the point of incidence. That is, the normal is along the radius, the line joining the centre of curvature of the mirror to the point of incidence.

10.2.1 Sign Convention

To derive the relevant formulae for reflection by mirrors (spherical) and refraction by lenses (Section 10.5), we must first adopt a sign convention. In this book, we shall follow the *New Cartesian sign convention*. According to this convention, all distances are measured from the pole of the mirror or the optical centre of the lens. The distances measured in the same direction as the incident light are taken as positive and those measured in the direction opposite to the direction of incident light are taken as negative. The heights measured upwards (above x -axis) and normal to the principal axis of the mirror/lens are taken as positive. The heights measured downwards are taken as negative.

What is the need for a sign convention for mirror and lens formulae? Why can we not write these formulae using only the magnitudes of distances of objects and images? We can do so, but then we shall get formulae with different

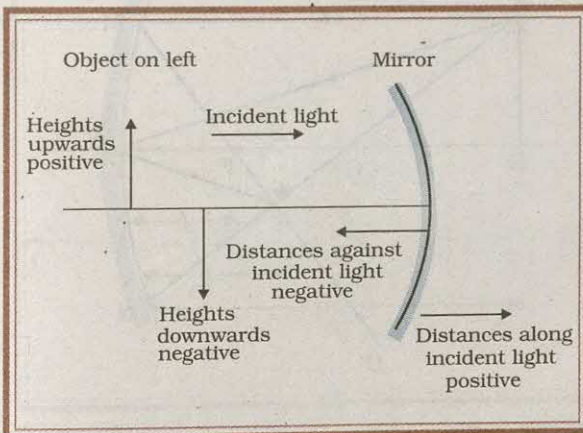


Fig. 10.1 The New Cartesian Sign Convention.

signs of terms for different cases. With a common accepted convention, it turns out that a single formula for spherical mirrors and a single formula for spherical lenses can handle all different cases. Care should be taken to substitute the numerical values with proper signs in these formulae.

10.2.2 Focal Length

Figure 10.2 shows what happens when a parallel beam of light is incident on (a) a concave mirror, and (b) a convex mirror. We assume that the rays are paraxial i.e., they are incident at points close to the pole of the mirror V . Also in the figure, we have taken the rays to be parallel to the axis CV of the mirror. The reflected rays converge at a point F on the axis for a concave mirror. For a convex mirror, the reflected rays appear to diverge from a point F . The point F is called the principal focus of the mirror. If the parallel paraxial (close to the principal axis) beam were incident making some angle with the axis, the reflected rays would converge (or appear to diverge) from a point in a plane through F normal to the axis. This is called the focal plane of the mirror.

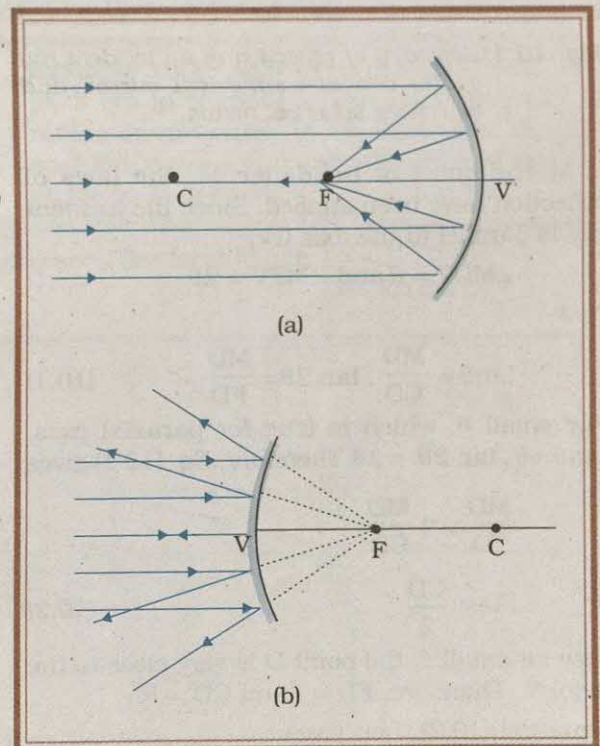


Fig. 10.2 Focus of a concave and convex mirror.

The distance between the focus F and the pole V of the mirror is called the focal length of the mirror, denoted by f . We now show that $f = R/2$, where R is the radius of curvature of the mirror. The geometry of reflection of an incident ray is shown in Fig. 10.3.

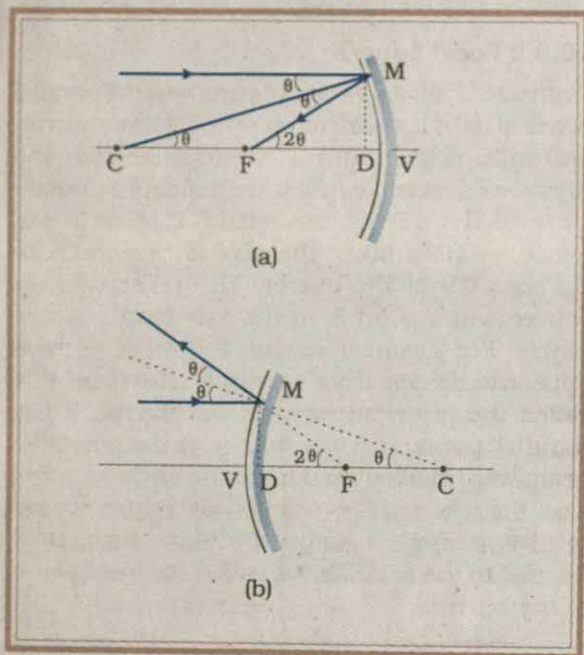


Fig. 10.3 Geometry of reflection of an incident ray on (a) concave spherical mirror, and (b) convex spherical mirror.

At the point of incidence M , the laws of reflection have been applied. Since the incident ray is parallel to the axis CV ,

$$\angle MCV = \theta \text{ and } \angle MFV = 2\theta$$

Now

$$\tan \theta = \frac{MD}{CD}, \tan 2\theta = \frac{MD}{FD} \quad (10.1)$$

For small θ , which is true for paraxial rays, $\tan \theta = \theta$, $\tan 2\theta = 2\theta$. Therefore, Eq. (10.1) gives

$$\frac{MD}{FD} = 2 \frac{MD}{CD}$$

$$\text{or } FD = \frac{CD}{2} \quad (10.2)$$

Now for small θ , the point D is very close to the point V . Therefore, $FD = f$ and $CD = R$

Equation (10.2) then gives

$$f = R/2 \quad (10.3)$$

10.2.3 The Mirror Equation

If rays starting from a point meet at another point after reflection and/or refraction, that point is called the image of the first point. The image is real if the rays actually converge to the point; it is virtual if the rays do not actually meet but appear to diverge from the point when produced backwards. Image is thus a point-to-point correspondence with the object established through reflection and/or refraction.

Considering reflection from a spherical mirror, in principle, we can take any two rays coming from an object, trace their paths, find their point of intersection and thus, obtain the image of the point. In practice, however, it is convenient to choose any two of the following rays:

- The ray from the point which is parallel to the principal axis. The reflected ray goes through the focus of the mirror.
- The ray passing through the centre of curvature of a concave mirror or appearing to pass through it for a convex mirror. The reflected ray simply retraces the path.
- The ray passing through the focus of the concave mirror or appearing to pass through (or directed towards) the focus of a convex mirror. The reflected ray is parallel to the principal axis.

With this method, Fig. 10.4 gives the ray diagram showing the image (in this case, real) of an object formed by a concave mirror.

We now derive the mirror equation or the relation between the object distance (u), image distance (v) and the focal length (f).

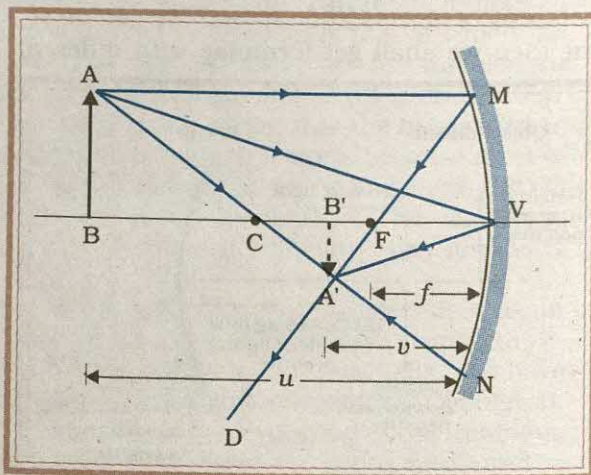


Fig. 10.4 The ray diagram for image formation by a concave mirror.

From the geometry of the figure, the two right-angled triangles $A'B'F$ and MVF are similar. (For paraxial rays, MV can be considered to be a straight line perpendicular to CV). Therefore,

$$\frac{A'B'}{MV} = \frac{B'F}{VF}$$

or

$$\frac{A'B'}{AB} = \frac{B'F}{VF} \quad (10.4)$$

The right angled triangles $A'B'V$ and ABV are also similar. Therefore,

$$\frac{A'B'}{AB} = \frac{B'V}{BV} \quad (10.5)$$

Comparing Eqs. (10.4) and (10.5), we get

$$\frac{B'V - VF}{VF} = \frac{B'V}{BV} \quad (10.6)$$

Equation (10.6) is a relation involving magnitudes of the distance. We now apply the sign convention:

$$B'V = -v, VF = -f, BV = -u \quad (10.7)$$

to get

$$\frac{-v + f}{-f} = \frac{-v}{-u}$$

or

$$\frac{v - f}{f} = \frac{v}{u}$$

$$\frac{1}{v} + \frac{1}{u} = \frac{1}{f} \quad (10.8)$$

This relation is known as the **mirror equation**.

The size of the image relative to the size of the object is another important quantity to consider. We define *magnification* (m) of a mirror as the ratio of the size of the image (h') to the size of the object (h):

$$m = \frac{h'}{h} \quad (10.9)$$

The size of the object h is always taken to be positive. Then h' is negative if the image is inverted, and positive if the image is erect. Since the triangles $A'B'V$ and ABV are similar, we have,

$$\frac{A'B'}{AB} = \frac{B'V}{BV}$$

With the sign convention,

$$\frac{-h'}{h} = \frac{-v}{-u}$$

so that

$$m = \frac{h'}{h} = -\frac{v}{u} \quad (10.10)$$

The mirror equation, Eq. (10.8), and the magnification formula, Eq. (10.10), though derived here for the case of real, inverted image by a concave mirror, are in fact valid for all cases of reflection by a spherical mirror (concave or convex) including virtual images. Figure 10.5 shows the ray diagrams for virtual image formation by a concave and convex mirror. You should verify that Eqs. (10.8) and (10.10) hold good for these cases as well.

Example 10.1 An object is placed (i) 10 cm, (ii) 5 cm in front of a concave mirror of radius of curvature 15 cm. Calculate the position, nature, and magnification of the image in each case.

Answer The focal length $f = -15/2$
 $= -7.5$ cm

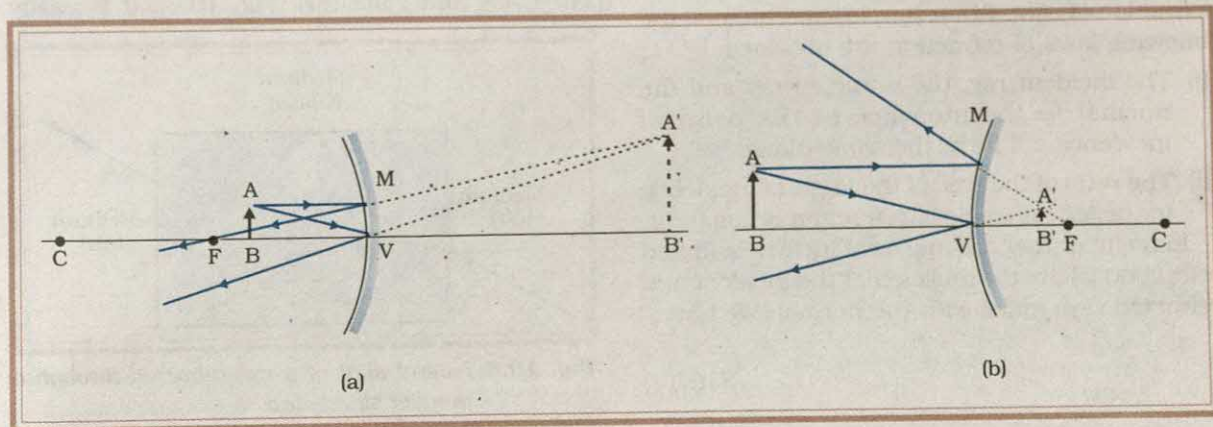


Fig. 10.5 The image formation by (a) concave mirror with object closer than F (b) convex mirror.

- (i) The object distance $u = -10$ cm.

$$\text{Since, } \frac{1}{u} + \frac{1}{v} = \frac{1}{f}$$

$$\frac{1}{v} + \frac{1}{-10} = \frac{1}{-7.5}$$

$$\text{or } v = \frac{10 \times 7.5}{7.5 - 10} = \frac{10 \times 7.5}{-2.5} = -30 \text{ cm}$$

The image is 30 cm from the mirror on the object side.

$$\text{Also, magnification } m = -\frac{v}{u} = -\frac{(-30)}{(-10)} = -3$$

The image is magnified, real and inverted.

- (ii) The object distance $u = -5$ cm

$$\text{Since } \frac{1}{u} + \frac{1}{v} = \frac{1}{f}$$

$$\frac{1}{v} + \frac{1}{-5} = \frac{1}{-7.5}$$

$$\text{or } v = \frac{5 \times 7.5}{7.5 - 5} = 15 \text{ cm}$$

This image is 15 cm at the back of the mirror (i.e., not on the object side). It is a virtual image.

$$\text{Magnification } m = -\frac{v}{u} = -\frac{15}{(-5)} = 3$$

The image is magnified, virtual and erect. ◀

10.3 REFRACTION

When light travels from one medium to another, it changes the direction of its path at the interface of the two media. This is called refraction of light. From simple experiments, the following laws of refraction are obtained.

- The incident ray, the refracted ray and the normal to the interface at the point of incidence, all lie in the same plane.
- The ratio of the sine of the angle of incidence to the sine of angle of refraction is constant.

Remember that the angles of incidence (i) and refraction (r) are the angles that the incident and refracted rays make with the normal. We have

$$\frac{\sin i}{\sin r} = n_{21} \quad (10.11)$$

where n_{21} is a constant, called the refractive index of the second medium with respect to the first medium. We note that n_{21} is a characteristic of the pair of media (and also depends on the wavelength of light*), but is independent of the angle of incidence. Equation (10.11) is the well-known Snell's law of refraction. We shall see in Chapter 11 that n_{21} is simply the ratio of the speed of light in medium 1 (v_1) to that in medium 2 (v_2).

$$n_{21} = \frac{v_1}{v_2}$$

From Eq. (10.11), if $n_{21} > 1$, $r < i$, i.e., the refracted ray bends towards the normal. The medium 2 is said to be optically denser (or denser for short) than medium 1. On the other hand, if $n_{21} < 1$, $r > i$, the refracted ray bends away from the normal. This is the case when incident ray in a denser medium refracts into a rarer medium. If the first medium is vacuum (or in

practice, air), the quantity $\frac{c}{v}$ (where v is the speed of light in the second medium) is called the absolute refractive index n (or refractive index for short) of the second medium. Clearly, then

$$n_{21} = \frac{n_2}{n_1}$$

This equation shows that

$$n_{12} = \frac{1}{n_{21}} \quad (10.12)$$

$$\text{and } n_{32} = n_{31} \times n_{12} \quad (10.13)$$

Some elementary results based on the laws of refraction follow immediately. For a rectangular slab, refraction takes place at two interfaces (air-glass and glass-air) (Fig. 10.6). It is easily

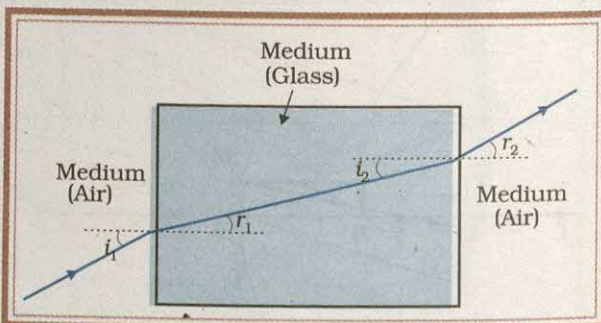


Fig. 10.6 Lateral shift of a ray refracted through a parallel sided slab.

* This is strictly valid only in case of a dispersive medium.

seen that $r_2 = i_1$ i.e., the emergent ray is parallel to the incident ray — there is no deviation, but it does suffer lateral displacement with respect to the incident ray. Another familiar observation is that the bottom of a tank or pond filled with water appears to be raised (Fig. 10.7). For viewing near the normal direction, it can be shown that the apparent depth is real depth divided by the refractive index of the medium (water).

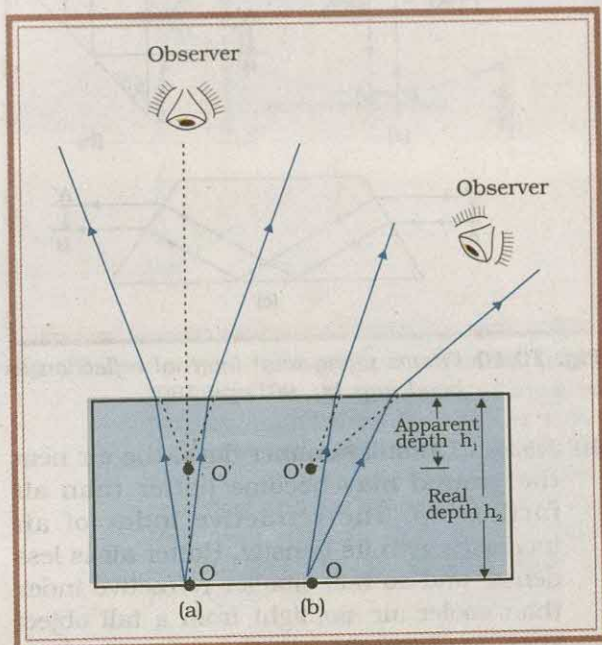


Fig. 10.7 Apparent depth for (a) normal and (b) oblique incidence.

The atmospheric refraction is responsible for many interesting phenomena. The Sun is visible a little before the actual sunrise and a little after the actual sunset (Fig. 10.8). By actual sunrise

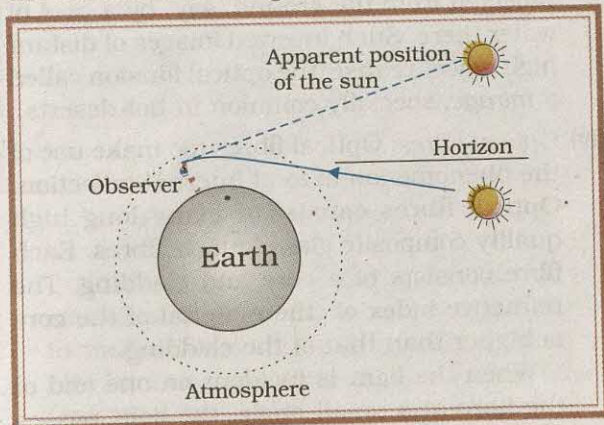


Fig. 10.8 Delayed sunset and advance sunrise due to atmospheric refraction.

we mean the actual crossing of the horizon by the Sun. We show in Fig. 10.8, the actual and apparent positions of the Sun with respect to the horizon. The figure is highly exaggerated to show the effect. Refractive index of air with respect to free or outer space is 1.00029. Due to this, the apparent shift in the direction of the Sun is by about half a degree ($1/2^\circ$) and the corresponding time difference between actual sunset and apparent sunset is about 2 min. The apparent flattening of the Sun at sunset and sunrise is also due to the same phenomenon.

10.4 TOTAL INTERNAL REFLECTION

When light passes from an optically denser medium to a rarer medium at the interface, it is partly reflected back into the same medium and partly refracted to the second medium. This reflection is called the *internal reflection*. Under certain conditions, the incident light can be made to be reflected back into the same medium without any significant loss of intensity. This gives rise to an interesting phenomenon known as *total internal reflection*.

When a ray of light travels from a denser medium to a rarer medium, the refracted ray is bent away from the normal, e.g., AO_1B in Fig. 10.9. The incident ray AO_1 is partially reflected (O_1C) and partially transmitted (O_1B) or refracted, the angle of refraction (r) being larger than the angle of incidence (i). As the angle of incidence increases, so does the angle of refraction, till for the ray AO_3 , the angle of refraction is 90° . The refracted ray is bent so much away from the normal that it grazes the surface and the interface between the two media. This is the ray AO_3D . If the angle of incidence is increased still further (e.g., the ray AO_4)

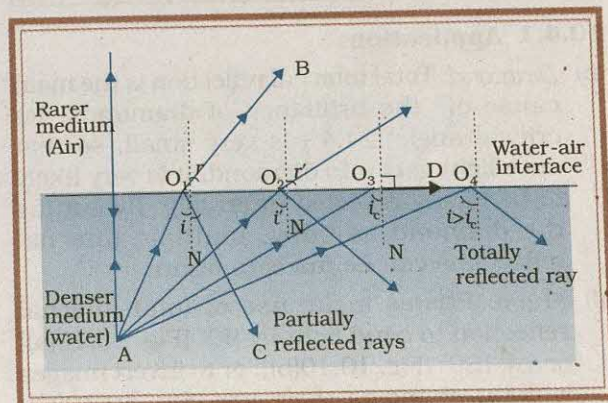


Fig. 10.9 Total internal reflection.

refraction is not possible, and the incident ray is totally reflected. This is called *total internal reflection*.

The critical angle of incidence, $\angle AO_3N$ is such that the angle of refraction is 90° . We see from Snell's law [Eq.(10.11)] that if the relative refractive index is less than one, then since the maximum value of $\sin r$ is unity, there is an upper limit to the value of $\sin i$ for which the law can be satisfied. This is $i = i_c$ such that

$$\sin i_c = n_{21}$$

For larger values of i , i.e., larger value of $\sin i$, Snell's law of refraction cannot be satisfied for any value of r (since the maximum value of $\sin r$ is unity). There is no refraction.

The critical angle for total internal reflection when light is incident on a rarer medium 2 from a denser medium 1 is given from Snell's law, by

$$\frac{\sin i_c}{\sin(\pi/2)} = \sin i_c = n_{21} = \frac{n_2}{n_1} \quad (10.14)$$

since by definition the angle of refraction for critical incidence is $(\pi/2)$ radians or 90° .

Some typical critical angles are listed in Table 10.1.

Table 10.1 Critical Angle for some Transparent Media

Substance	Refractive index	Critical angle
Water	1.33	48.75°
Crown glass	1.52	41.14°
Dense flint glass	1.62	37.31°
Diamond	2.42	24.41°

10.4.1 Applications

- (i) **Diamond:** Total internal reflection is the main cause of the brilliance of diamonds. Its critical angle (24.4°) is very small, so that once light gets into diamond, it is very likely to be totally reflected internally. By cutting the diamond suitably, multiple internal reflections can be made to occur.
- (ii) **Prism:** Prisms make use of total internal reflection to bend light by 90° [Fig. 10.10(a)] or by 180° [Fig. 10.10(b)], or to invert images without changing their size [Fig. 10.10(c)]. In the first two cases, the critical angle i_c for

the material of the prism must be less than 45° . We see from Table 10.1 that this is true for both crown glass and dense flint glass.

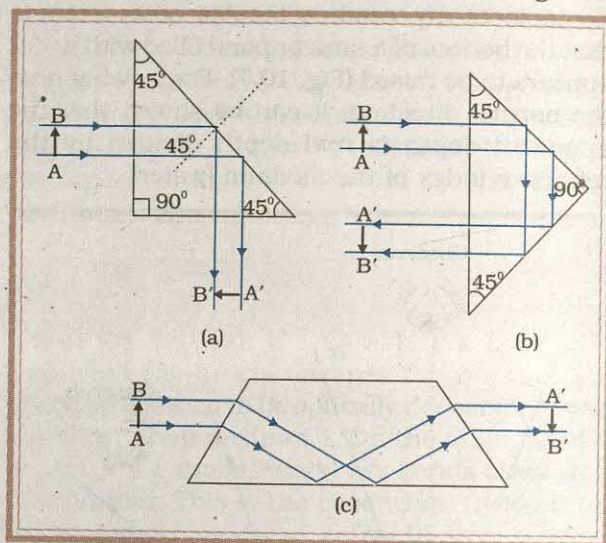


Fig. 10.10 Prisms using total internal reflection to bend rays by 90° and 180° .

(iii) **Mirage:** On still summer days, the air near the ground may become hotter than air further up. The refractive index of air increases with its density. Hotter air is less dense, and so has smaller refractive index than cooler air. So, light from a tall object such as a tree passes through a medium whose refractive index decreases towards the ground. Thus a ray of light from such an object gets bent and is totally internally reflected. This is shown in Fig. 10.11(b). This light reaches the observer in the direction shown. He naturally assumes that it is reflected from the ground, say, by a pool of water there. Such inverted images of distant high objects cause the optical illusion called a *mirage*, specially common in hot deserts.

(iv) **Optical fibres:** Optical fibres too make use of the phenomenon of total internal reflection. Optical fibres consist of many long high quality composite glass/quartz fibres. Each fibre consists of a core and cladding. The refractive index of the material of the core is higher than that of the cladding.

When the light is incident on one end of the fibre at a small angle, the light passes inside, undergoes repeated total internal reflections along the fibre and finally comes

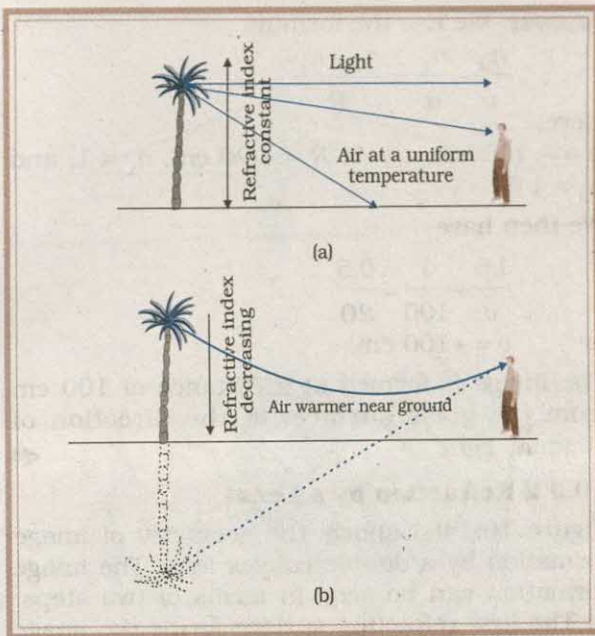


Fig. 10.11 (a) A tree is seen by an observer when the air above the ground is at uniform temperature. (b) When the air close to the ground is hot, light bends gradually as shown, undergoing total internal reflection, and the apparent image of the tree may mislead the observer into thinking that there is a pool of water in front of the tree!

out (Fig. 10.12). The angle of incidence is always larger than the critical angle of the core material with respect to its cladding. Even if the fibre is bent, the light can easily travel through along the fibre.

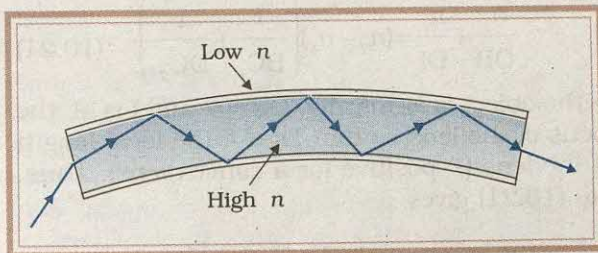


Fig. 10.12 An optical fibre.

A bundle of optical fibres can be put to several uses. It can be used as a 'light pipe' in medical and optical examination. It can also be used for optical signal transmission. Optical fibres have also been used for transmitting and receiving electrical signals which are converted to light by suitable

transducers. The main requirement is that there should be very little absorption of light as it travels for long distances inside the optical fibre. This has been achieved by purification and special preparation of materials such as quartz. In silica glass fibres, it is possible to transmit more than 95% of the light over a fibre length of 1 km. (Compare with what you expect for block of ordinary window glass 1 km thick). (For more details see Chapter 16.)

10.5 REFRACTION AT SPHERICAL SURFACES AND BY LENSES

We have so far considered refraction at a plane interface. Any small part of a spherical surface can be regarded as planar and the same laws of refraction can be applied at every point on the surface. Just as for reflection by a spherical mirror, the normal at the point of incidence is perpendicular to the tangent plane to the surface at that point and, therefore, passes through the centre of curvature of the surface. We first consider refraction by a single spherical surface. A thin lens is a transparent optical medium bounded by two spherical surfaces. Applying the formula for image formation by a single spherical surface successively at the two surfaces of a lens, we obtain the thin lens formula and the lens maker's formula.

10.5.1 Refraction at a Spherical Surface

Figure 10.13 shows the geometry of formation of image I of an object point O on the principal axis of the spherical surface with centre of curvature C , and radius of curvature R . The rays are incident from a medium of refractive index n_1 , to another of refractive index n_2 . As before,

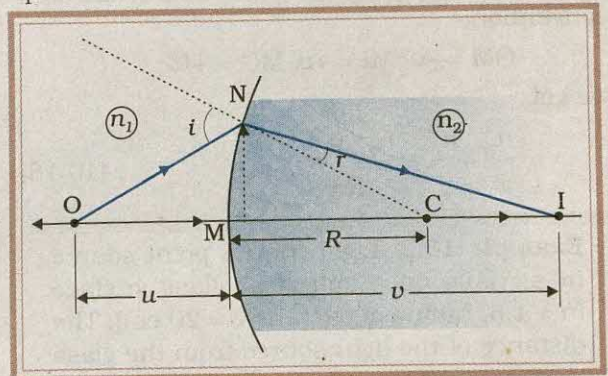


Fig. 10.13 Refraction at a spherical surface separating two media.

we take the aperture (or the lateral size) of the surface to be small compared to other distances involved, so that small angle approximation can be made, wherever appropriate. In particular, NM will be taken to be nearly equal to the length of the perpendicular from N to the principal axis. We have

$$\tan \angle NOM = \angle NOM = \frac{MN}{OM} \quad (\text{for small angles}),$$

$$\tan \angle NCM = \angle NCM = \frac{MN}{MC} \quad (\text{for small angles}),$$

$$\tan \angle NIM = \angle NIM = \frac{MN}{MI} \quad (\text{for small angles})$$

Now, for $\triangle NOC$, i is the exterior angle. Therefore,
 $i = \angle NOM + \angle NCM$

$$= \frac{MN}{OM} + \frac{MN}{MC} \quad (10.15)$$

Similarly,

$$r = \angle NCM - \angle NIM$$

$$\text{i.e., } r = \frac{MN}{MC} - \frac{MN}{MI} \quad (10.16)$$

Now, by Snell's law

$$n_1 \sin i = n_2 \sin r$$

or for small angles

$$n_1 i = n_2 r$$

Substituting i and r from Eqs. (10.15) and (10.16), we get

$$\frac{n_1}{OM} + \frac{n_2}{MI} = \frac{n_2 - n_1}{MC} \quad (10.17)$$

Here, OM, MI, MC represent magnitudes of distances. Applying the New Cartesian sign convention,

$$OM = -u, MI = +v, MC = +R.$$

we get

$$\frac{n_2}{v} - \frac{n_1}{u} = \frac{n_2 - n_1}{R} \quad (10.18)$$

Answer We use the formula

$$\frac{n_2}{v} - \frac{n_1}{u} = \frac{n_2 - n_1}{R}$$


Here,

$$u = -100 \text{ cm}, v = ?, R = +20 \text{ cm}, n_1 = 1, \text{ and } n_2 = 1.5$$

We then have

$$\frac{1.5}{v} + \frac{1}{100} = \frac{0.5}{20}$$

$$\text{or } v = +100 \text{ cm.}$$

The image is formed at a distance of 100 cm from the glass surface, in the direction of incident light. 

10.5.2 Refraction by a Lens

Figure 10.14(a) shows the geometry of image formation by a double convex lens. The image formation can be seen in terms of two steps: (i) The first refracting surface forms the image I_1 of the object O. [Fig. 10.14(b)]. The image I_1 acts as a virtual object for formation of image I by the second surface [Fig. 10.14(c)]. Applying Eq. (10.17) to the first interface ABC, we get:

$$\frac{n_1}{OB} + \frac{n_2}{BI_1} = \frac{n_2 - n_1}{BC_1} \quad (10.19)$$

A similar procedure applied to the second interface ADC gives,

$$-\frac{n_2}{DI_1} + \frac{n_1}{DI} = \frac{n_1 - n_2}{DC_2} \quad (10.20)$$

For a thin lens, $BI_1 = DI_1$. Adding Eqs. (10.19) and (10.20), we get:

$$\frac{n_1}{OB} + \frac{n_1}{DI} = (n_2 - n_1) \left(\frac{1}{BC_1} + \frac{1}{DC_2} \right) \quad (10.21)$$

If the object is at infinity, $OB = \infty$ and I is at the focus of the lens so that $DI = f$, the focal length of the lens (f positive for a convex lens). Thus, Eq. (10.21) gives

$$\frac{n_1}{f} = (n_2 - n_1) \left(\frac{1}{BC_1} + \frac{1}{DC_2} \right) \quad (10.22)$$

By the sign convention,

$$BC_1 = +R_1, DC_2 = -R_2$$

so that Eq. (10.22) gives:

$$\frac{1}{f} = (n_2 - n_1) \left(\frac{1}{R_1} - \frac{1}{R_2} \right) \quad (10.23)$$

Example 10.2 Light from a point source in air falls on a spherical glass surface ($n = 1.5$, radius of curvature = 20 cm). The distance of the light source from the glass surface is 100 cm. At what position is the image formed?

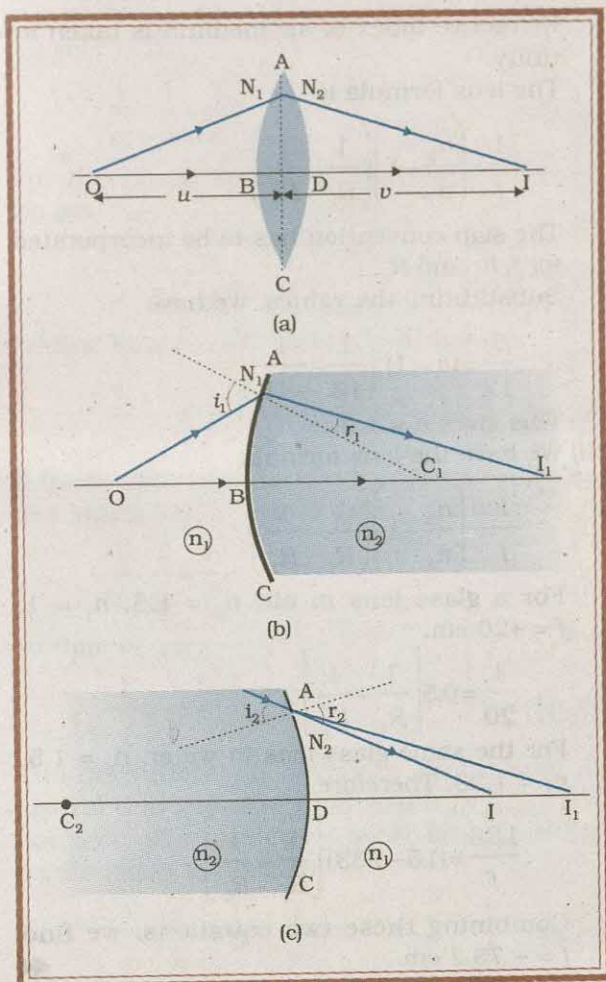


Fig. 10.14 Thin lens formula. (a) The object, the double convex lens and the image. (b) Refraction at the first spherical surface. (c) Refraction at the second spherical surface.

Equation (10.23) is known as the **lens maker's formula**. It is evidently, useful to design lenses of desired focal length using surfaces of suitable radii of curvature. Note that the formula is true for a concave lens also. In that case R_1 is negative, R_2 positive and therefore, f is negative. From Eqs. (10.21) and (10.22), we get

$$\frac{n_1}{OB} + \frac{n_1}{DI} = \frac{n_1}{f} \quad (10.24)$$

Again, in the thin lens approximation, B and D are both close to the optical centre of the lens. Applying the sign convention, $OB = -u$, $DI = +v$, we get

$$\frac{1}{v} - \frac{1}{u} = \frac{1}{f} \quad (10.25)$$

Equation (10.25) is the familiar **thin lens formula**. Though we derived it for a convex lens, the formula is valid for both convex as well as concave lenses and for both real and virtual images.

To find the image of an object by a lens, we can, in principle, take any two rays coming from an object point and trace their paths using the laws of refraction and find the point where the refracted rays meet (or appear to meet). In practice, however, it is convenient to choose any two of the following rays:

- (i) A ray from the object parallel to the principal axis of the lens after refraction passes through the second principal focus F' (in a convex lens) or appears to diverge (in a concave lens) from the first principal focus F .
- (ii) A ray of light, passing through the optical centre of the lens, emerges without any deviation after refraction.
- (iii) A ray of light passing through the first principal focus (for a convex lens) or appearing to meet at it (for a concave lens) emerges parallel to the principal axis after refraction.

Figures 10.15(a) and (b) illustrate these rules for a convex and a concave lens. You should

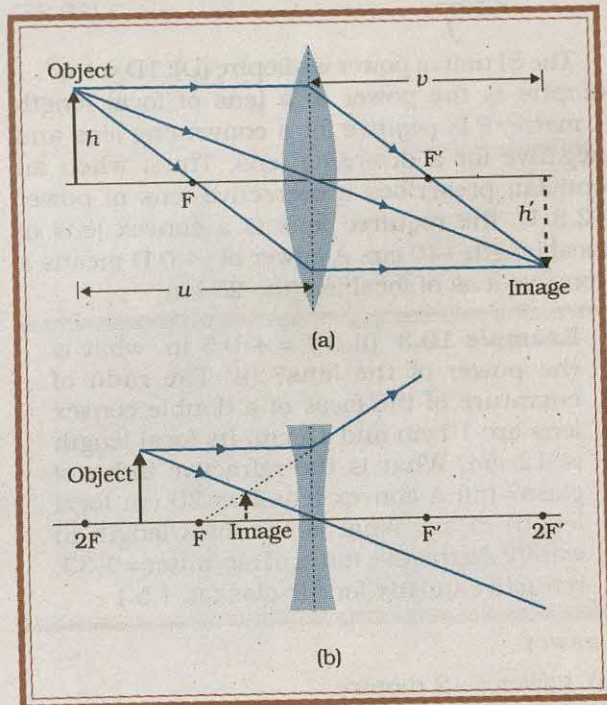


Fig. 10.15 Tracing rays through (a) convex lens (b) concave lens.

practice drawing similar ray diagrams for different position of the object from the lens and also verify that the lens formula, Eq. (10.25), holds good for all cases.

Magnification (m) produced by a lens is defined, like that for a mirror, as the ratio of the size of the image to that of the object. The size of the object h is always taken to be positive, but image size h' is positive for erect image and negative for an inverted image. Proceeding in the same way as for spherical mirrors, it is easily seen that for a lens

$$m = \frac{h'}{h} = + \frac{v}{u} \quad (10.26)$$

Thus, for erect (and virtual) image formed by a convex or concave lens, m is positive, while for an inverted (and real) image, m is negative.

10.5.3 Power of a Lens

Power of a lens is a measure of the additional convergence or divergence which a lens introduces in the light falling on it. Clearly, a lens of shorter focal length bends the incident light more, refracted light converging in case of convex lens and diverging in case of concave lens. The power of a lens (P) is defined to be the reciprocal of its focal length as

$$P = \frac{1}{f} \quad (10.27)$$

The SI unit of power is diopetre (D): $1\text{D} = 1\text{m}^{-1}$. 1 diopetre is the power of a lens of focal length 1 metre. P is positive for a converging lens and negative for a diverging lens. Thus, when an optician prescribes a corrective lens of power +2.5 D, the required lens is a convex lens of focal length +40 cm. A power of -4.0 D means a concave lens of focal length -25 cm.

Example 10.3 (i) If $f = +0.5\text{ m}$, what is the power of the lens? (ii) The radii of curvature of the faces of a double convex lens are 10 cm and 15 cm. Its focal length is 12 cm. What is the refractive index of glass? (iii) A convex lens has 20 cm focal length in air. What is the focal length in water? (Refractive index of air-water = 1.33, refractive index for air-glass is 1.5.)

Answer

- (i) Power = +2 diopetre.
 (ii) Here, we have $f = +12\text{ cm}$, $R_1 = +10\text{ cm}$, $R_2 = -15\text{ cm}$.

Refractive index of air medium is taken as unity.

The lens formula is

$$\frac{1}{f} = \left(\frac{n_2}{n_1} - 1 \right) \left(\frac{1}{R_1} - \frac{1}{R_2} \right)$$

The sign convention has to be incorporated for f , R_1 and R_2 .

Substituting the values, we have

$$\frac{1}{12} = (n - 1) \left(\frac{1}{10} - \frac{1}{-15} \right)$$

This gives $n = 1.5$.

(iii) We have the lens formula

$$\frac{1}{f} = \left(\frac{n_2}{n_1} - 1 \right) \left(\frac{1}{R_1} - \frac{1}{R_2} \right)$$

For a glass lens in air, $n_2 = 1.5$, $n_1 = 1$, $f = +20\text{ cm}$.

$$\frac{1}{20} = 0.5 \left[\frac{1}{R_1} - \frac{1}{R_2} \right]$$

For the same glass lens in water, $n_2 = 1.5$, $n_1 = 1.33$. Therefore,

$$\frac{1.33}{f} = (1.5 - 1.33) \left[\frac{1}{R_1} - \frac{1}{R_2} \right]$$

Combining these two equations, we find $f = +78.2\text{ cm}$.

10.5.4 Combination of Thin Lenses in Contact

Consider two lenses A and B of focal length f_1 and f_2 placed in contact with each other. Let the object be placed at O beyond the focus of the first lens A (Fig. 10.16). The first lens produces an image at I_1 . This serves as a virtual object for the second lens B, producing the final image at I. Since the lenses are thin, we take the optical centres of the lenses to be coincident. Let this central point be denoted by P.

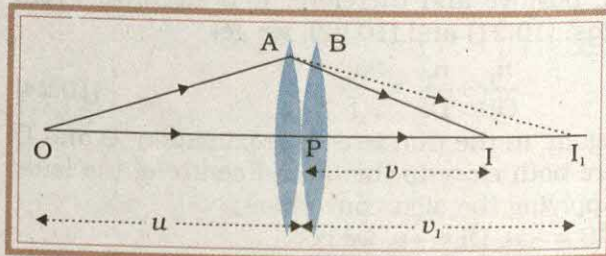


Fig. 10.16 Image formation by two thin lenses in contact.

For the image formed by the first lens A, we get

$$\frac{1}{v_1} - \frac{1}{u} = \frac{1}{f_1} \quad (10.28)$$

For the image formed by the second lens B, we get

$$\frac{1}{v} - \frac{1}{v_1} = \frac{1}{f_2} \quad (10.29)$$

Adding Eqs. (10.28) and (10.29), we get

$$\frac{1}{v} - \frac{1}{u} = \frac{1}{f_1} + \frac{1}{f_2} \quad (10.30)$$

If the two lens-system is regarded as equivalent to a single lens of focal length f , we have

$$\frac{1}{v} - \frac{1}{u} = \frac{1}{f}$$

so that we get

$$\frac{1}{f} = \frac{1}{f_1} + \frac{1}{f_2} \quad (10.31)$$

The derivation will obviously go through for several thin lenses of focal length f_1, f_2, f_3, \dots , in contact. The effective focal length of the combination is given by

$$\frac{1}{f} = \frac{1}{f_1} + \frac{1}{f_2} + \frac{1}{f_3} + \dots \quad (10.32)$$

In terms of power, Eq. (10.32) can be written as

$$P = P_1 + P_2 + P_3 + \dots \quad (10.33)$$

where P is the net power of the lens combination. Note that the sum in Eq. (10.33) is an algebraic sum of individual powers i.e., some of the terms on the right side may be positive (for convex lenses) and some negative (for concave lenses). Lens combination helps to increase the magnification and sharpness of the image. Since the image formed by the first lens becomes the object for the second, Eq. (10.26) implies that the total magnification m of the combination is a product of magnification (m_1, m_2, m_3, \dots) of individual lenses:

$$m = m_1 \times m_2 \times m_3 \times \dots \quad (10.34)$$

Such a system of combination of lenses is commonly used in the design of objectives of cameras, microscopes, telescopes and other optical instruments.

Example 10.4 Find the position of the image formed by the lens combination given in the Fig. 10.17 (a) below:

Answer Image formed by the first lens

$$\frac{1}{v_1} - \frac{1}{u_1} = \frac{1}{f_1}$$

$$\frac{1}{v_1} - \frac{1}{-30} = \frac{1}{10}$$

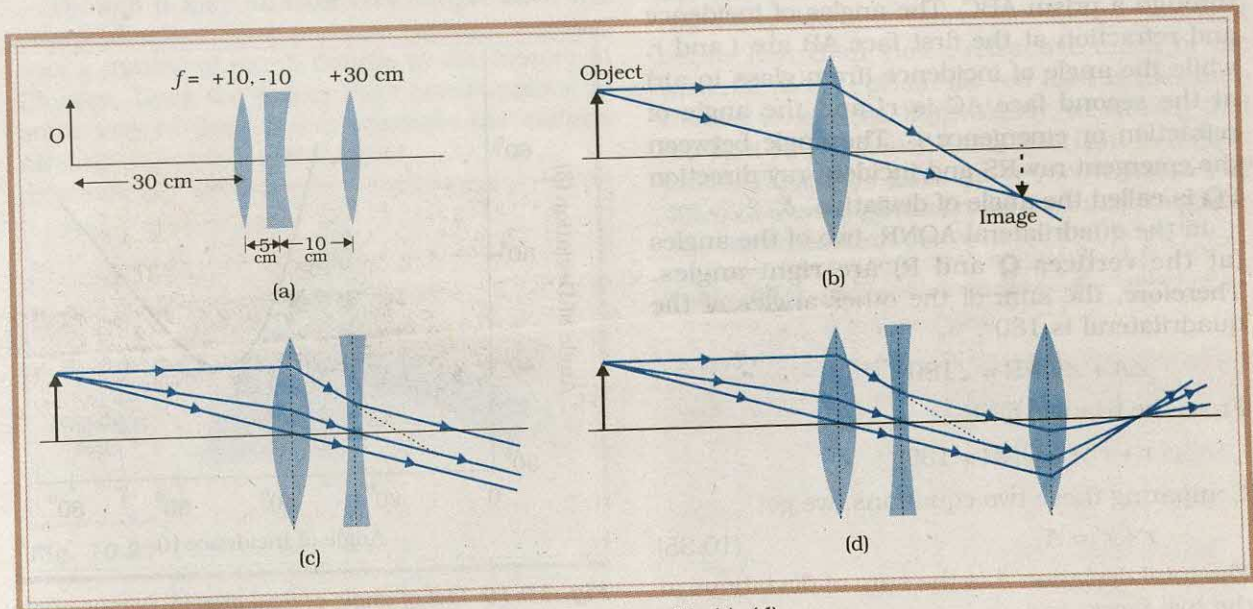


Fig. 10.17 (a), (b), (c), (d).

$$\text{or } v_1 = 15 \text{ cm}$$

The image formed by the first lens serves as the object for the second. This is at a distance of $(15 - 5) \text{ cm} = 10 \text{ cm}$ to the right of the second lens. It is a virtual object.

$$\frac{1}{v_2} - \frac{1}{10} = \frac{1}{-10}$$

$$\text{or } v_2 = \infty$$

The virtual image is formed at an infinite distance to the left of the second lens. This acts as an object for the third lens.

$$\frac{1}{v_3} - \frac{1}{u_3} = \frac{1}{f_3}$$

$$\text{or } \frac{1}{v_3} = \frac{1}{\infty} + \frac{1}{30}$$

$$\text{or } v_3 = 30 \text{ cm.}$$

The final image is formed 30 cm to the right of the third lens.

Ray diagrams and image formation for the above combination of lenses are given in the Fig. 10.17(b) - (d). Figure shows the ray diagram and image formation, for the first lens [Fig. 10.17(b)]; for the first and second lens [Fig. 10.17(c)] and finally for the combination of all the three lenses [Fig. 10.17(d)].

10.6 REFRACTION IN A PRISM

Figure 10.18 shows the passage of a light through a prism ABC. The angles of incidence and refraction at the first face AB are i and r , while the angle of incidence (from glass to air) at the second face AC is r' and the angle of refraction or emergence i' . The angle between the emergent ray RS and incident ray direction PQ is called the angle of deviation, δ .

In the quadrilateral AQNR, two of the angles (at the vertices Q and R) are right angles. Therefore, the sum of the other angles of the quadrilateral is 180° .

$$\angle A + \angle QNR = 180^\circ$$

From the triangle QNR,

$$r + r' + \angle QNR = 180^\circ$$

Comparing these two equations, we get

$$r + r' = A \quad (10.35)$$

The total deviation δ is the sum of deviations at the two faces:

$$\delta = (i - r) + (i' - r')$$

$$\text{i.e. } \delta = i + i' - A \quad (10.36)$$

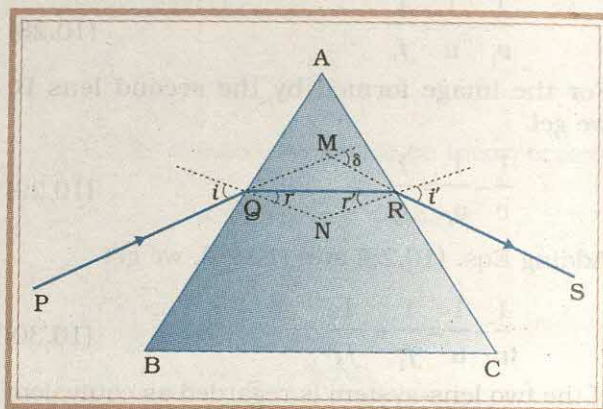


Fig. 10.18 Light ray passing through a triangular glass prism.

A plot between the angle of deviation and angle of incidence is shown in Fig. 10.19. You can see that, in general, any given value of δ corresponds to two values i and i' . This fact is expected from the symmetry of i and i' in Eq. (10.36) i.e., δ remains the same if i and i' are interchanged. Physically, this is related to the fact that the path of ray in Fig. 10.18 can be traced back, resulting in the same angle of deviation. At the minimum deviation

$\delta = D_m$, $i = i'$ which implies $r = r'$. Eq. (10.35) gives

$$2r = A \text{ or } r = \frac{A}{2} \quad (10.37)$$

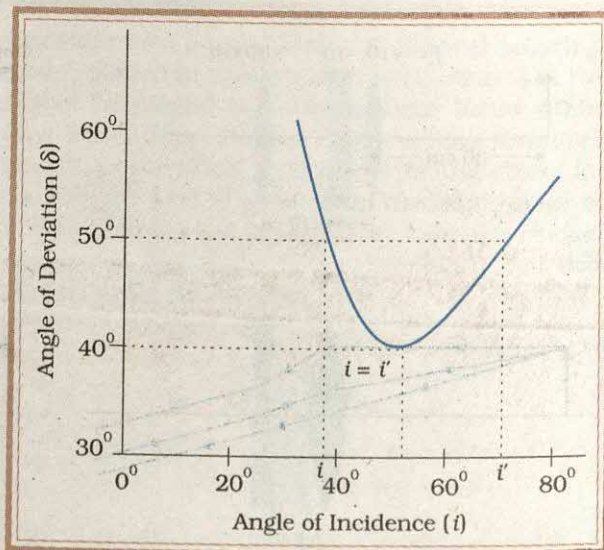


Fig. 10.19 Plot of angle of deviation (δ) versus angle of incidence (i) for a triangular prism.

In the same way, Eq. (10.36) gives

$$D_m = 2i - A, \text{ or } i = (A + D_m)/2 \quad (10.38)$$

The refractive index of the prism (or generally, the refractive index of the material of the prism with respect to the medium outside) is:

$$n_{21} = \frac{n_2}{n_1} = \frac{\sin[(A + D_m)/2]}{\sin[A/2]} \quad (10.39)$$

The angles A and D_m can be measured experimentally with good precision. Equation (10.39) thus provides an accurate method for measuring refractive index of the prism.

10.7 DISPERSION BY A PRISM

Dispersion is the splitting of light into its component colours and the pattern of colour components of light is called its spectrum. The word 'spectrum' is now used in a much more general sense: we discussed in Chapter 9 the electromagnetic spectrum over the large range of wavelengths, from γ -rays to radiowaves, of which the spectrum of light (visible spectrum) is only a small part.

When a narrow beam of sunlight is incident on a glass prism, the emergent light is seen to be consisting of several colours. There is actually a continuous variation of colour, but broadly, the different component colours in sequence are: violet, indigo, blue, green, yellow, orange and red (given by the acronym VIBGYOR). The red light bends the least, while the violet light bends the most.

Though it may all look very simple now, the origin of colour after passage through a prism was a matter of much debate in the history of Physics. Does the prism itself create colour in some way or does it only separate the colours already present in white light?

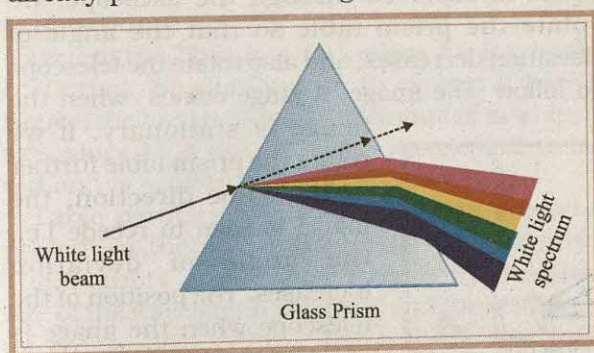


Fig. 10.20 Dispersion of sunlight or white light passing through a glass prism. The relative deviation of different colours is shown highly exaggerated.

In a classic experiment known for its simplicity but great significance, Isaac Newton settled the issue once for all. He put another similar prism, but in an inverted position, and let the emergent separated beam from the first prism fall on the second prism Fig. 10.21. The resulting emergent beam was found to be white light. The explanation was clear; the first prism separated the white light into its component colours, which then were recombined by the inverted prism to give white light, thus white light itself consists of colours which are separated by the prism.

We now know that colour is associated with

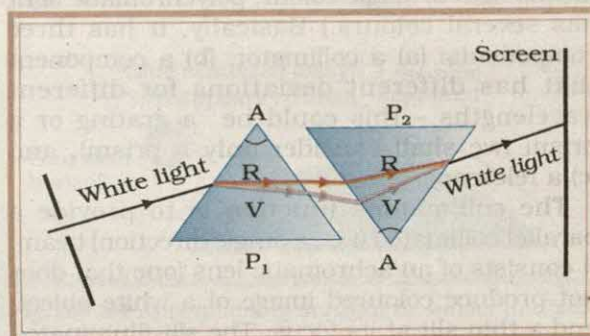


Fig. 10.21 Schematic diagram of Newton's classic experiment on dispersion of white light.

wavelength of light. In the visible spectrum, red light is at the long wavelength end (~ 700 nm) while the violet light is at the short wavelength end (~ 400 nm). Dispersion takes place because the refractive index of medium for different colours is different. Since, for example, red light bends less than violet, refractive index of the material of the prism for red light is less than that for violet light. Equivalently, we can say that red light travels faster than violet light in a glass prism. Table 10.2 gives the refractive indices for different wavelength for crown glass and flint glass. The variation of refractive index with wavelength

Table 10.2 Refractive Indices for different Wavelengths

Colour	Wavelength (nm)	Crown glass	Flint glass
Violet	396.9	1.533	1.663
Blue	486.1	1.523	1.639
Yellow	589.3	1.517	1.627
Red	656.3	1.515	1.622

The variation of refractive index with wavelength may be more pronounced in some media than the other. In vacuum, of course, the speed of light is independent of wavelength. Thus vacuum (or air approximately) is a non-dispersive medium in which all colours travel with the same speed. On the other hand, glass is a dispersive medium.

10.8 SPECTROMETER

A spectrometer shown in Fig. 10.22 is an optical instrument to produce the spectrum and measure different wavelengths from the source of polychromatic light. (Monochromatic light means light of single colour, polychromatic light has several colours.) Basically, it has three components: (a) a collimator, (b) a component that has different deviations for different wavelengths – this could be a grating or a prism (we shall consider only a prism), and (c) a telescope.

The collimator's function is to provide a parallel collimated (i.e., a single direction) beam. It consists of an achromatic lens (one that does not produce coloured image of a white object) and a thin slit at its focus. The slit illuminated by the source of light behind it acts as a source so that the emergent rays from the lens are parallel. The parallel rays from the lens fall on the prism and get dispersed. The dispersed beam enters the telescope.

The objective lens of the telescope forms a real image of the slit in a different direction for each wavelength of light present. There may be discrete lines or a continuous band. The spectrum is magnified by an eye-piece. The prism is mounted on a rotating platform, and the collimator and telescope also can be rotated about a common axis perpendicular to the prism table. The positions of the collimator and telescope can be read off from a graduated circular vernier scale. In this way, the angle of deviation can be measured for each wavelength.

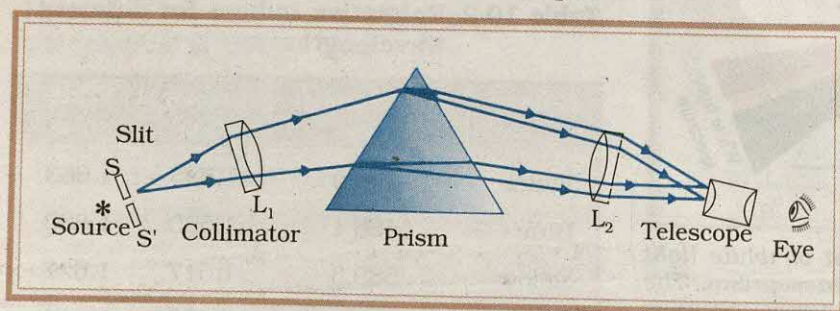


Fig. 10.22 Spectrometer.

10.8.1 Determination of Refractive Index

For the actual measurement, some adjustments are necessary. First, the collimator should be focussed for parallel rays. The telescope is adjusted as follows. The eye-piece of the telescope is adjusted so that the cross-wires are clearly visible. Next, the telescope is focussed for infinity. By removing the parallax between the image and the cross-wire, the telescope is set for parallel rays. Finally, the height of the prism table is so adjusted that the centre of the slit, prism and the telescope all lie in the same horizontal plane, and the image of the slit appears in the centre of the field of view.

To determine refractive index, according to Eq. (10.39), we need to know the angle of the prism A and the angle of minimum deviation D_m .

To determine A , the prism is so placed that the two faces of the prism making the angle A face the collimator directly. That is, incident parallel light from the collimator falls on each face. The telescope is first brought to view the image of the slit formed, by reflection from the first face, and the reading of its position on the vernier is noted. Next, it is rotated so as to view the image of the slit formed by reflection from the second face, and the reading is noted again. The difference of two readings is θ , which can be shown geometrically to be $2A$. Thus, the angle of prism $A = \theta/2$ is determined.

To determine the angle of minimum deviation D_m , we first view the image of the slit directly through the telescope (without placing the prism) and note its reading.

Next, the prism is placed on the table and the image of the slit after refraction through the prism is observed through the telescope. We rotate the prism table so that the angle of deviation decreases, and also rotate the telescope to follow the image. A stage comes when the image is stationary. If we rotate the prism table further in the same direction, the image is seen to recede i.e., the angle of deviation increases. The position of the telescope when the image is stationary is noted. The difference between the two readings of the angular position of the telescope gives

the angle of minimum deviation D_m . This procedure is repeated for different wavelengths. Using Eq. (10.39), and substituting the measured values of A and D_m (for different wavelengths), we can determine the refractive index of the material of the prism for different wavelengths.

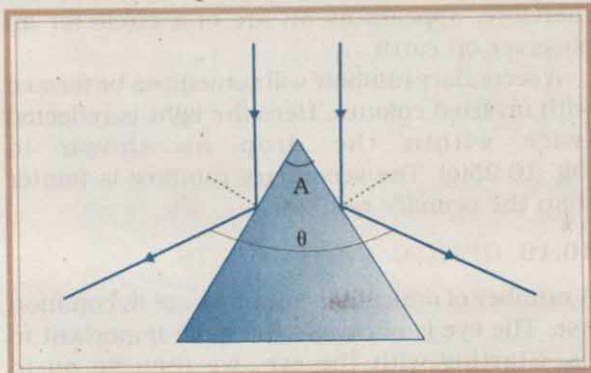


Fig. 10.23 Determining angle of the prism.

10.9 LIGHT IN NATURE

The shenanigans (interplay) of light with things around us gives rise to several beautiful phenomena. The blue of the sky, white clouds, the redhues of sunrise and sunset, the rainbow, the brilliant colours of some pearls, shells, and wings of birds, are just a few of the natural wonders we are used to. We describe some of them here from the point of view of Physics.

10.9.1 Scattering of Light

As sunlight travels through the earth's atmosphere, it gets *scattered* (changed in direction) by the large number of molecules present. Light of shorter wavelengths is scattered much more than light of longer wavelengths. (The amount of scattering is inversely proportional to the fourth power of the wavelength. This law is known as Rayleigh scattering. We shall not try to explain it here.) Hence, the bluish colour predominates in a clear sky, since blue has a shorter wavelength than red and is scattered much more strongly.

Large particles like dust and water droplets present in the atmosphere behave differently. The quantity of relevance here is the relative size of the wavelength of light λ , and the scatterer (of typical size say a). For $a \ll \lambda$, one has Rayleigh scattering which is proportional to $(1/\lambda)^4$. For $a \gg \lambda$, i.e., large scattering objects (e.g., raindrops, large dust or ice particles) this is not true; all wavelengths are scattered nearly

equally. Thus, clouds which have droplets of water with $a \gg \lambda$ are generally white.

At sunset or sunrise, the Sun's rays must pass through a larger atmospheric distance (Fig. 10.24). Most of the blue and shorter wavelengths are removed by scattering. The unscattered light which the eye receives therefore looks redder. This explains the appearance of the Sun and full moon near the horizon. These may look almost reddish.

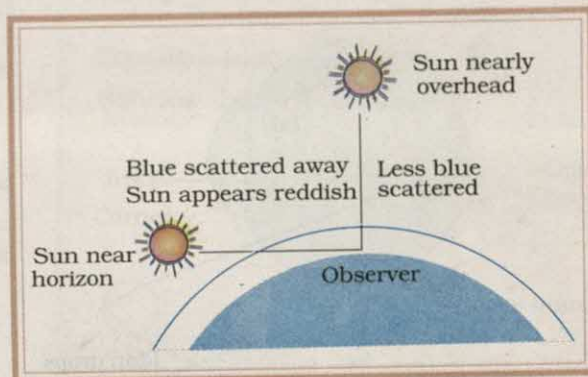


Fig. 10.24 Absorption of sunlight at sunset and sunrise.

10.9.2 The Rainbow

The rainbow is another example of the dispersion of sunlight by the water drops in the atmosphere [Fig. 10.25(a)]. This is a phenomenon due to a combination of the refraction of sunlight by spherical water droplets and of internal (not total) reflection. Fig. 10.25(b) shows how the sunlight is broken into its segments in the process and a rainbow appears. The dispersion of the violet and the red rays in the drop is shown in the Fig. 10.25(b).

Because of the spherical shape of the drop, even though the incident beam is parallel, many values of the angle of incidence occur for rays meeting the drop at different points. Therefore, unlike in the case of a prism, the emerging light is not a parallel beam. It is, however, found that the intensity of the red light is maximum at an angle of 42° and that of the violet rays at another angle (40°). The reason for this intensity peak is as follows. Rays which are once reflected have a maximum deviation of 180° (Ray travelling along a diameter and back). The minimum deviation occurs for a particular angle of incidence i . As Fig. 10.25(b) shows, there is a very slow variation of δ near the minimum. So rays with a large range of i are compressed into a small range of

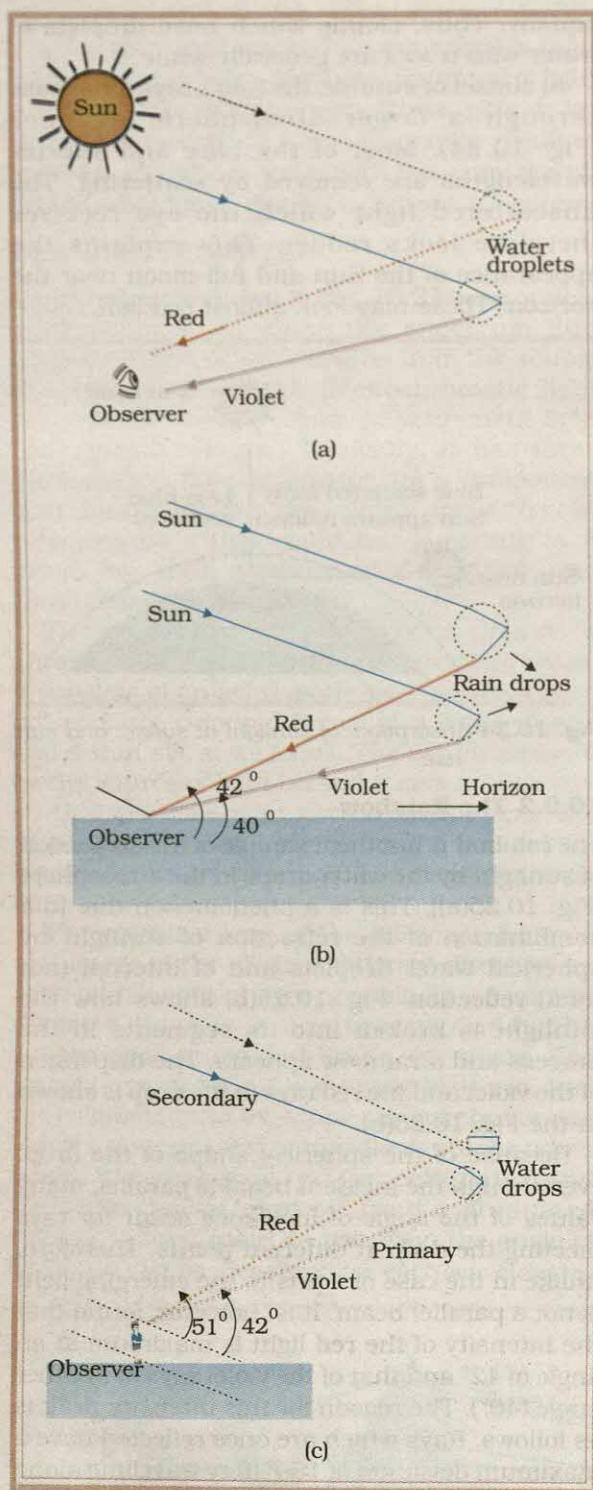


Fig. 10.25 Rainbow: (a) The Sun, water drops, and the eye; (b) Internal reflection and refraction of red and violet rays from raindrops, and (c) Primary and secondary rainbows.

δ , giving high intensity. Rays undergoing two internal partial reflections have a *maximum* deviation of about 50° for red and 54° for violet. The parallel beam of sunlight getting dispersed at these angles produces a cone of rays at the observer as seen in the figure. The rainbow, therefore, appears as an arc of a circle for an observer on earth.

A secondary rainbow will sometimes be formed with inverted colours. Here the light is reflected twice within the drop as shown in Fig. 10.25(c). The secondary rainbow is fainter than the primary rainbow.

10.10 OPTICAL INSTRUMENTS

A number of optical instruments are in common use. The eye is of course the most important to us. Starting with the eye, we then go on to describe the principles of working of the camera, the microscope and the telescope.

10.10.1 The Eye

Figure 10.26(a) shows the eye. Light enters the eye through a curved front surface, the cornea. It passes through the pupil which is the central hole in the iris. The size of the pupil can change under control of muscles. The light is further focussed by the eye-lens on the retina. The retina is a film of nerve fibres covering the curved back surface of the eye. The retina contains *rods* and *cones* which sense light intensity and colour, respectively, and transmit electrical signals via the optic nerve to the brain which finally processes this information. The shape (curvature) and therefore the focal length of the lens can be modified somewhat by the *ciliary muscles*. For example, when the muscle is relaxed, the focal length is about 2.5 cm and (for a normal eye) objects at infinity are in sharp focus on the retina. When the object is brought closer to the eye, in order to maintain the same image-lens distance (≈ 2.5 cm), the focal length of the eye-lens becomes shorter by the action of the ciliary muscles. This property of the eye is called *accommodation*. If the object is too close to the eye, the lens cannot curve enough to focus the image on to the retina, and the image is blurred. The closest distance for which the lens can focus light on the retina is called the *least distance of distinct vision*, or the *near point*. The standard value (for normal vision) taken here is 25 cm. (Often the near point is given the symbol D .)

This distance increases with age, because of the decreasing effectiveness of the ciliary muscle and the loss of flexibility of the lens. The near point may be as close as about 7 to 8 cm in a child ten years of age, and may increase to as much as 200 cm at 60 years of age. Thus, if an elderly person tries to read, placing the book at about 25 cm from the eye, the image appears blurred. This condition (defect of the eye) is called *presbyopia*. It is corrected by using a converging lens for reading.

Example 10.5 What focal length should the reading spectacles have for a person whose D value is 50 cm?

Answer The book is at $u = -25$ cm, the image should be formed at $v = -50$ cm. Therefore, the desired focal length is given by

$$\frac{1}{f} = \frac{1}{v} - \frac{1}{u}$$

$$\text{or } \frac{1}{f} = \frac{1}{-50} - \frac{1}{-25} = \frac{1}{50}$$

$$\text{or } f = +50 \text{ cm (convex lens).}$$

A number of optical defects of the eye are common. For example, the lens may converge incident light to a point well before the retina. This is called *nearsightedness* or *myopia*. This means that the eye is producing too much convergence in the incident beam. To compensate this, we interpose a concave lens between the eye and the object, with the right diverging effect. The image will now be focussed on the retina Fig. 10.26(b).

Similarly, if the lens focuses at a point behind the retina, a convergent lens is needed to compensate. This defect is called *farsightedness* or *hypermetropia* Fig. 10.26(c).

Another common defect of vision is called *astigmatism*. This occurs when the cornea is not spherical in shape. For example, the cornea could have a larger curvature in the vertical plane than in the horizontal plane. If one looks at a horizontal wire or line, focussing in the vertical plane is needed for a sharp image. Astigmatism results in lines in one direction being well focussed while those in a perpendicular direction are not. It is corrected

by a lens with one surface which is of cylindrical rather than spherical shape. A cylindrical surface focuses rays in one plane but not in a perpendicular plane. By choosing the radius of curvature and axis direction of the cylindrical surface, astigmatism can be corrected. This can occur along with myopia or hypermetropia.

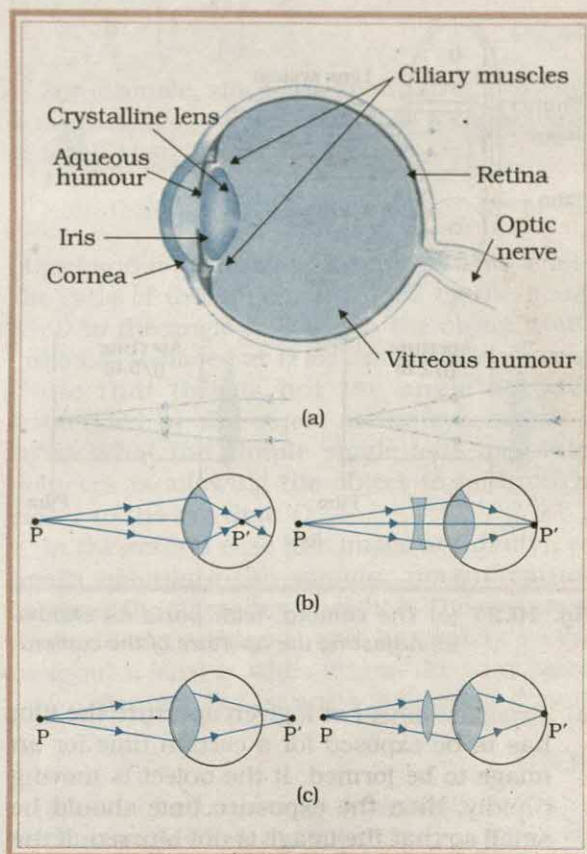


Fig. 10.26 (a) The eye, showing some important parts; (b) the shortsighted, or myopic eye, and (c) The farsighted, or hypermetropic eye.

10.10.2 The Camera

A photographic camera consists of a converging lens system at one end of a light proof box and light sensitive film at the other end [Fig. 10.27(a)]. A real inverted image of the object is formed on the film. The total amount of light falling on the film can be adjusted by changing the aperture (diameter) of the lens.

A shutter is placed between the lens and the film. When a photograph is taken, the shutter opens and closes quickly, thus exposing the film for a short time to light entering the camera through the lens. The exposure time is another

parameter controlling the total light falling on the film.

For a good picture to be formed, the film should be exposed sufficiently but not too much. The following are some important aspects of photography.

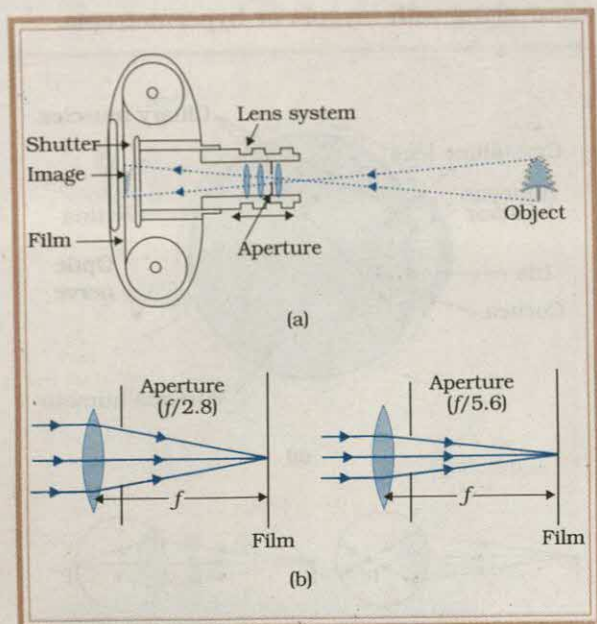


Fig. 10.27 (a) The camera, with parts as shown. (b) Adjusting the aperture of the camera.

- (a) **Exposure time:** For a given aperture the film has to be exposed for a certain time for an image to be formed. If the object is moving rapidly, then the exposure time should be small so that the image is not blurred. If the photograph is taken in bright sunlight, the exposure time can be small. For an object in the shade, or for indoor photography, the exposure time has to be larger. The normal exposure times used in cameras are

$$\frac{1}{500} \text{ s}, \frac{1}{250} \text{ s}, \frac{1}{125} \text{ s}, \frac{1}{60} \text{ s}, \frac{1}{30} \text{ s}, \text{ etc.}$$

- (b) **Aperture of the lens:** This refers to the diameter of the circular opening through which the light passes into the camera. This diameter is normally expressed as a fraction of f and is called the f -number. Some apertures used in a camera are $(f/2)$, $(f/2.8)$, $(f/4)$, $(f/5.6)$, $(f/8)$, $(f/11)$, $(f/16)$, etc. An aperture of $(f/2.8)$ means the diameter d of the aperture is $f/2.8$. The area corresponding

to aperture $f/5.6$ is one-fourth of the area corresponding to aperture $f/2.8$. In order to have the same exposure of the film, the shutter will have to be kept open for four times the previous duration. You may also note that the numbers 2, 2.8, 4, 5.6, 8, 11, 16 are obtained by a successive multiplication by a factor of $\sqrt{2}$ i.e., these apertures correspond to shutter times (for getting the same exposure) which differ successively by a factor of $(\sqrt{2})^2 = 2$.

- (c) **Film speed:** The film speed is a measure of how quickly the film will be exposed when in use. For the same light level a 'fast' film needs a relatively short time exposure while a 'slow' speed film needs somewhat longer time. Fast films are, therefore, used in poor lighting conditions and very slow films for still object photography. (Slow films, if they can be used, have the advantage of giving sharper images.)
- (d) **Exposure meter:** Many cameras have built-in exposure meters. Such a meter has a light sensitive surface. Depending on the amount of light falling on it, a proportional amount of current flows in the meter. Using this, the photographer can adjust a suitable combination of aperture and exposure times for correct exposure. In many cameras this happens automatically.
- (e) **Depth of focus:** Strictly speaking, only one object distance u corresponds to perfect focus on the film. The depth of focus indicates the range Δu of the object distances around u over which the focusing is reasonably good. You can see from Fig. 10.27(b) that increasing the aperture reduces the depth of focus.

It will be interesting to practice and understand the various details of photography described above. In addition there are many further aids like wide angle and telephoto lenses, close-up adaptors which make photography an exciting hobby.

10.10.3 The Microscope

A simple magnifier or microscope is a converging lens of short focal length (Fig. 10.28). The lens is held near the object, one focal length away or less, and the eye is positioned close to the lens on the other side. The idea is to form an erect,

magnified and virtual image of the object at a distance so that the eye can view it comfortably, i.e., 25 cm or greater. If the object is at a distance f , the image is at infinity. If the object is at a smaller distance, the image is virtual and closer than infinity. The closest comfortable distance for viewing is the near point (distance $D \approx 25$ cm). Viewing at the near point causes some strain on the eye, so that often the image formed is at infinity, suitable for viewing by the relaxed eye. We show both cases, the first in Fig. 10.28(a), and the second in Figs. 10.28(b) and (c).

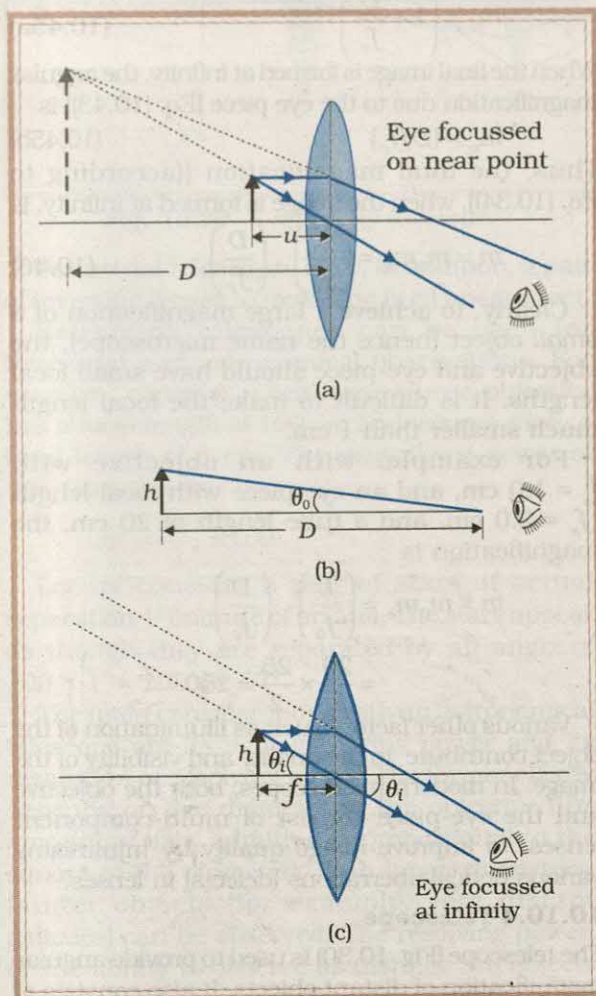


Fig. 10.28 A simple microscope. (a) The magnifying lens is located such that the image is at the near point. (b) The object by itself, at the near point. (c) The object at the focal point of the lens; the image is at infinity.

The magnification when the image is formed at the near point D is easily calculated. We have

the linear magnification m to be

$$m = \frac{v}{u} = v \left(\frac{1}{v} - \frac{1}{f} \right) = \left(1 - \frac{v}{f} \right)$$

Now according to our sign convention, v is negative, and is equal in magnitude to D . Thus the magnification is

$$m = \left(1 + \frac{D}{f} \right) \quad (10.40)$$

For example, since D is about 25 cm, to have a magnification of six, one needs a convex lens of focal length $f = 5$ cm.

Note that $m = \frac{h'}{h}$ where h is the size of the

object and h' the size of the image. This is also the ratio of the angle subtended by the image h'/D to the angle h/D which the object would subtend, if placed at D for comfortable viewing. (Note that this is not the angle actually subtended by the object at the eye, which is h/u .) What the simple single lens magnifier achieves is allowing the object to be brought closer to the eye than D .

In the second case (the image at infinity), we again calculate the angular magnification. Suppose the object has a height h . The maximum angle it can subtend, and be clearly visible (without a lens) is when it is at the near point, i.e., a distance D . The angle subtended then is

$$\theta_0 = \left(\frac{h}{D} \right) \quad (10.41)$$

We now find the angle subtended at the eye by the image when the object is at u . From the relations

$\frac{h'}{h} = m = \frac{v}{u}$ we have the angle subtended by the image

$\theta_i = \frac{h'}{-v} = \frac{h}{-v} \cdot \frac{v}{u} = \frac{h}{-u}$ = angle subtended by the object, which is now at $u = -f$.

$$\theta_i = \left(\frac{h}{f} \right) \quad (10.42)$$

as is clear from Fig. 10.28(c). The angular magnification is therefore

$$m = \left(\frac{\theta_i}{\theta_0} \right) = \frac{D}{f} \quad (10.43)$$

This is one less than the magnification when the image is at the near point, Eq. (10.40), but the viewing is more comfortable, and the difference in magnification usually small. In subsequent discussions of optical instruments (microscope and telescope) we shall assume the image to be at infinity.

A simple microscope has a limited maximum magnification (≤ 10) for realistic focal lengths. For much larger magnifications, one uses two lenses, one compounding (enhancing) the effect of the other. This is known as the *compound microscope*. A schematic diagram is shown in Fig. 10.29. The lens nearest the object, called the objective, forms a real, inverted, magnified image of the object. This serves as the object for the second lens, the eye-piece, which functions essentially like a simple microscope or magnifier, producing an enlarged virtual final image. The first inverted image is thus near (at or within) the focal point of the eye-piece, at a distance appropriate for final image formation at infinity, or a little closer for image formation at the near point. Clearly, the final image is inverted with respect to the original object.

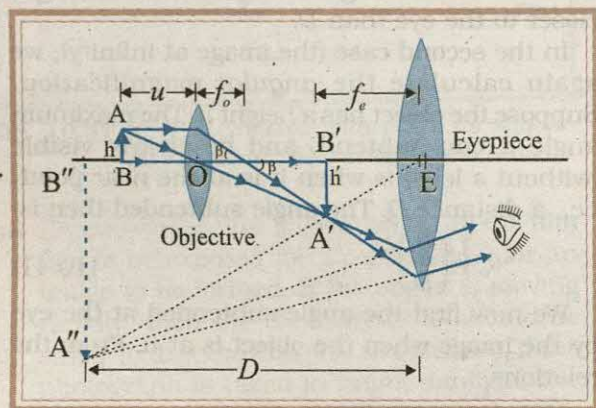


Fig. 10.29 A compound microscope.

We now obtain the magnification due to the compound microscope. The ray diagram of Fig. 10.29 shows that the (linear) magnification due to the objective, namely, (h'/h) , equals

$$m_o = \frac{h'}{h} = \frac{L}{f_o} \quad (10.44)$$

where we have used the result

$$\tan \beta = \left(\frac{h}{f_o} \right) = \left(\frac{h'}{L} \right)$$

Here h' is the size of the first image, the object size being h and the objective focal length being

f_o . The first image is formed near the focal point of the eye-piece. The distance L , i.e., the distance between the second focal point of the objective and the first focal point of the eyepiece (focal length f_e) is called the tube length of the compound microscope, since both f_o and f_e are rather small.

As the inverted first image is near the focal point of the eye-piece, we use the result from the discussion above for the simple microscope that the (angular) magnification m_e due to it [Eq. (10.40)], when the final image is formed at the near point, is

$$m_e = \left(1 + \frac{D}{f_e} \right) \quad (10.45a)$$

When the final image is formed at infinity, the angular magnification due to the eye piece [Eq. (10.43)] is

$$m_e = (D/f_e) \quad (10.45b)$$

Thus, the total magnification [(according to Eq. (10.34)], when the image is formed at infinity, is

$$m = m_o m_e = \left(\frac{L}{f_o} \right) \left(\frac{D}{f_e} \right) \quad (10.46)$$

Clearly, to achieve a large magnification of a small object (hence the name microscope), the objective and eye-piece should have small focal lengths. It is difficult to make the focal length much smaller than 1 cm.

For example, with an objective with $f_o = 1.0$ cm, and an eye-piece with focal length $f_e = 2.0$ cm, and a tube length of 20 cm, the magnification is

$$\begin{aligned} m &= m_o m_e = \left(\frac{L}{f_o} \right) \left(\frac{D}{f_e} \right) \\ &= \frac{20}{1} \times \frac{25}{2} = 250 \end{aligned}$$

Various other factors such as illumination of the object contribute to the quality and visibility of the image. In modern microscopes, both the objective and the eye-piece consist of multi-component lenses, to improve image quality by minimising various optical aberrations (defects) in lenses.

10.10.4 Telescope

The telescope (Fig. 10.30) is used to provide angular magnification of distant objects. It also consists of an objective and an eye-piece. But here, the objective has a large focal length and a much larger aperture than the eye piece. Light from a distant object enters the objective and a real image is formed in the tube at the second focal point of the convex objective lens. The eye-piece magnifies this image producing a final inverted image. The magnifying power m corresponds to angular

magnification. Angular magnification is defined as the ratio of the angle β subtended at the eye by the image to the angle α which the object subtends at the lens or the eye. Hence

$$m = \frac{\beta}{\alpha} \approx \frac{A'B'}{f_e} \cdot \frac{f_o}{AB} = \frac{f_o}{f_e} \quad (10.47)$$

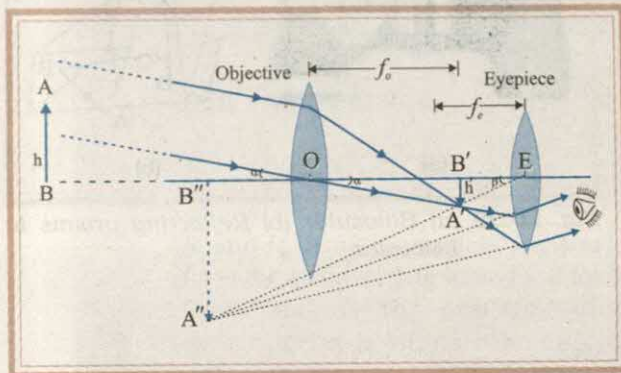


Fig. 10.30 A refracting telescope.

Terrestrial telescopes have, in addition, a pair of inverting lenses to make the final image erect.

Refracting telescopes can be used for terrestrial and astronomical observations. For example, consider a telescope whose objective has a focal length of 100 cm and the eye-piece a focal length of 1 cm. The magnifying power of this telescope is

$$m = \frac{100}{1} = 100$$

Let us consider a pair of stars of actual separation $1'$ (minute of an arc). The stars appear as though they are separated by an angle of $100 \times 1' = 100' = 1.67^\circ$.

The main considerations with an astronomical telescope are its light gathering power and its resolution or resolving power. The former clearly depends on the diameter of the objective (the amount of light admitted is proportional to the square of the diameter). With larger diameters, fainter objects (for example, very distant galaxies) can be observed. The resolving power, or the ability to observe as distinct two objects which are in very nearly the same direction, also depends on the diameter of the objective, as we shall see in the next Chapter. So, the desirable aim in optical telescopes is to make them of large objective diameter. The largest lens objective in use (refracting telescope objective) has a diameter of 102 cm (It is at the Yerkes Observatory in Wisconsin, USA). Such big lenses

tend to be very heavy and therefore, difficult to make and support by their edges.

For these reasons, modern telescopes use a concave mirror rather than a lens for the objective.

Telescopes with mirror objectives are called *reflecting telescopes*. They have several advantages. First, there is no chromatic aberration (formation of coloured image of white objects) in a mirror. Second, if a parabolic surface is chosen, spherical aberration (formation of non-point, blurred image of point objects) is also removed. Mechanical support is much less of a problem since the mirror weighs much less than a lens of equivalent optical quality, and can be supported over its entire back surface, not just over its rim. One obvious problem, with a reflecting telescope is that the objective mirror focuses light inside the telescope tube. One must have an eye-piece and the observer right there, obstructing some light (depending on the size of the observer cage). This is what is done in the very large 200 inch (~6.6 m) diameter Mt. Palomar telescope, California. The viewer sits near the focal point of the mirror, in a small cage. Another solution to the problem is to deflect the light being focussed by another mirror. One such arrangement, using a convex secondary mirror to focus the incident light, which now passes through a hole in the objective primary mirror, is shown in Fig. 10.31. This is known as a *Cassegrain telescope*, after its inventor. It has a large focal length in a short telescope. There are also other types of reflecting arrangements such as the Newtonian reflector and the Schmidt reflector. The largest telescope in India is in Kavalur, Tamil Nadu. It is a 2.34 m diameter

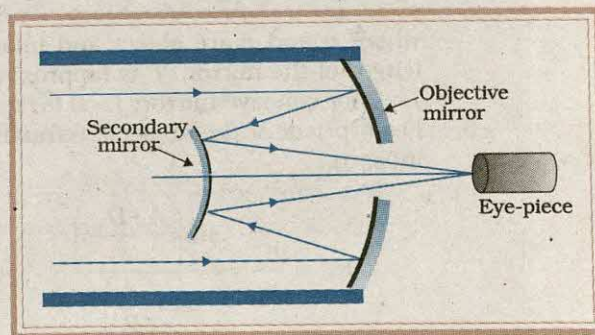


Fig. 10.31 A reflecting telescope (Cassegrain).

reflecting telescope (Cassegrain). It was ground, polished, set up and is being used by the Indian Institute of Astrophysics, Bangalore. The largest reflecting telescopes in the world are the pair of Keck telescopes in Hawaii, USA, of 10 metres in diameter.

The *prism binocular* is a double telescope using two sets of totally reflecting prisms. This makes the final image erect which is very desirable for observations on earth. It also effectively folds the optical path ABCDEF making for a shorter distance between the objective and the eye-piece (Fig. 10.32). Binoculars are much more compact and easier to use than a refracting telescope, and as the name indicates, allow use of both eyes.

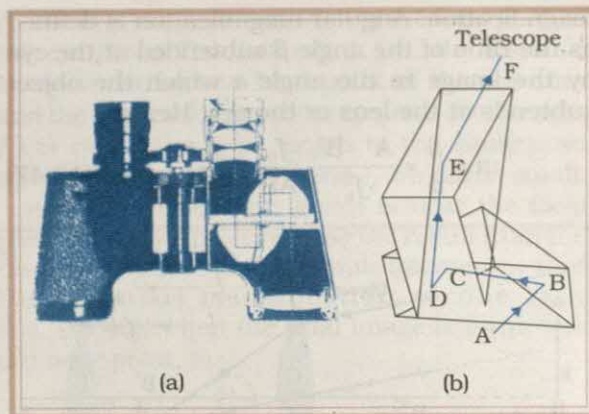


Fig. 10.32 (a) Binocular (b) Reflecting prisms in binocular.

SUMMARY

1. Reflection is governed by the equation $i = r$ and refraction by Snell's law, $\frac{\sin i}{\sin r} = n$, where the incident ray, reflected ray, refracted ray and normal all lie in the same plane. Angles of incidence, reflection and refraction are i , r and r' , respectively.
2. The *critical angle of incidence* i_c from a denser to rarer medium is given by

$\sin i_c = \frac{n_{\text{rarer}}}{n_{\text{denser}}}$. For $i > i_c$, total internal reflection occurs. Examples: multiple internal reflections in diamond ($i_c = 24.41^\circ$), totally reflecting prisms, mirage, etc.

Optical fibres consist of glass fibres coated with a thin layer of material of lower refractive index. Light incident at one end comes out after multiple internal reflections, even if the fibre is bent.

3. **New Cartesian sign convention:** Distances measured in the same direction as the incident light are positive; those measured in the opposite direction are negative. All distances are measured from the centre of the mirror/lens on the optic axis. The heights measured upwards above x -axis and normal to the principal axis of the mirror/lens are taken as positive. The heights measured downwards are taken as negative.
4. **Mirror equation:**

$$\frac{1}{v} + \frac{1}{u} = \frac{1}{f}$$

where u and v are object and image distances, respectively and f is the focal length of the mirror. f is (approximately) half the radius of curvature R . $f < 0$ for concave mirror; $f > 0$ for a convex mirror.

5. For a prism of the angle A , refractive index n_2 placed in a medium of refractive index n_1 ,

$$n_{21} = \frac{n_2}{n_1} = \frac{\sin\left(\frac{A + D_m}{2}\right)}{\sin\left(\frac{A}{2}\right)}$$

where D_m is the angle of minimum deviation.

6. *Refraction through a spherical interface* (from medium 1 to 2 of refractive index n_1 and n_2 , respectively)

$$\frac{n_2}{v} - \frac{n_1}{u} = \frac{n_2 - n_1}{R}$$

Thin lens formula

$$\frac{1}{v} - \frac{1}{u} = \frac{1}{f}$$

Lens maker's formula

$$\frac{1}{f} = \frac{(n_2 - n_1)}{n_1} \left(\frac{1}{R_1} - \frac{1}{R_2} \right)$$

R_1 and R_2 being the radii of curvature of the lens surfaces.

$f > 0$ for a converging lens; $f < 0$ for a diverging lens.

7. *Dispersion* is the splitting of light into its constituent colours. A spectrometer is an instrument to disperse polychromatic light and measure different wavelengths. It consists of a collimator, a prism and a telescope. The rainbow is an example of dispersion of sunlight by the water drops in the atmosphere.
8. *The Eye*: The eye has a convex length of focal length about 2.5 cm. This focal length can be varied somewhat so that the image is always on the retina. This ability of the eye is called *accommodation*. If the image is focussed before the retina (myopia), a diverging corrective lens is needed; if the image is focussed beyond the retina (hypermetropia), a converging corrective lens is needed.
9. In a *camera*, the aperture size and exposure time together determine the total amount of light received. An f -number of n means the diameter of the aperture is f/n . The sequence $f/2, f/2.8, f/4, f/5.6, f/8, f/11$ etc. is such that the exposure times required to get the same amount of light are in the ratio 1:2:4:8:16:32.
10. *Magnifying power of a simple microscope* has a magnitude m given by $m = 1 + (D/f)$, where $D = 25$ cm is the least distance of distinct vision and $f (> 0)$ is the focal length of the convex lens. If the image is at infinity, $m = D/f$.

For a compound microscope, the magnifying power is given by $m = m_e \times m_o$ where $m_e = 1 + (D/f_e)$ is the magnification due to the eye-piece and m_o is the magnification produced by the objective. *Approximately*,

$$m = \frac{L}{f_o} \times \frac{D}{f_e};$$

where f_o and f_e are the focal lengths of the objective and eye-piece respectively, and L is the distance between them.

11. *Magnifying power m of a telescope* is the ratio of the angle β subtended at the eye by the image to the angle α subtended at the eye by the object.

$$m = \frac{\beta}{\alpha} = \frac{f_o}{f_e},$$

where f_o and f_e are the focal lengths of the objective and eye-piece, respectively.

POINTS TO PONDER

1. The real image of an object placed between f and $2f$ from a convex lens can be seen on a screen placed at the image location. If the screen is removed, is the image still there? This question puzzles many, because it is difficult to reconcile ourselves with an image suspended in air without a screen. But the image does exist. Rays from a given point on the object are converging to an image point in space and diverging away. The screen simply diffuses these rays some of which reach our eye and we see the image.
2. Image formation needs regular reflection/refraction. In principle, all rays from a given point should reach the same image point. This is why you do not see your image by an irregular, reflecting object, say the page of a book.
3. For a simple microscope, the angular size of the object equals the angular size of the image. Yet it offers magnification because we can keep the small object much closer to the eye than 25 cm and hence have it subtend a large angle. The image is at 25 cm which we can see. Without the microscope, you would need to keep the small object at 25 cm which would subtend a very small angle.

EXERCISES

- 10.1 To print a photograph from a negative, the time of exposure to light from a lamp placed 0.50 m away is 2.5 s. How much exposure time is required if the lamp is placed 1.0 m away?
- 10.2 A small candle 2.5 cm in size is placed 27 cm in front of a concave mirror of radius of curvature 36 cm. At what distance from the mirror should a screen be placed in order to receive a sharp image? Describe the nature and size of the image. If the candle is moved closer to the mirror, how would the screen have to be moved?
- 10.3 A 4.5 cm needle is placed 12 cm away from a convex mirror of focal length 15 cm. Give the location of the image and the magnification. Describe what happens as the needle is moved farther from the mirror.
- 10.4 A square wire of side 3.0 cm is placed 25 cm away from a concave mirror of focal length 10 cm. What is the area enclosed by the image of the wire? (The centre of the wire is on the axis of the mirror, with its two sides normal to the axis.)
- 10.5 A tank is filled with water to a height of 12.5 cm. The apparent depth of a needle lying at the bottom of the tank is measured by a microscope to be 9.4 cm. What is the refractive index of water? If water is replaced by a liquid of refractive index 1.63 upto the same height, by what distance would the microscope have to be moved to focus on the needle again?
- 10.6 Figures 10.33 (a) and (b) show refraction of an incident ray in air at 60° with the normal to a glass-air and water-air interface, respectively. Predict the angle of refraction of an incident ray in water at 45° with the normal to a water-glass interface [Fig. 10.33(c)].

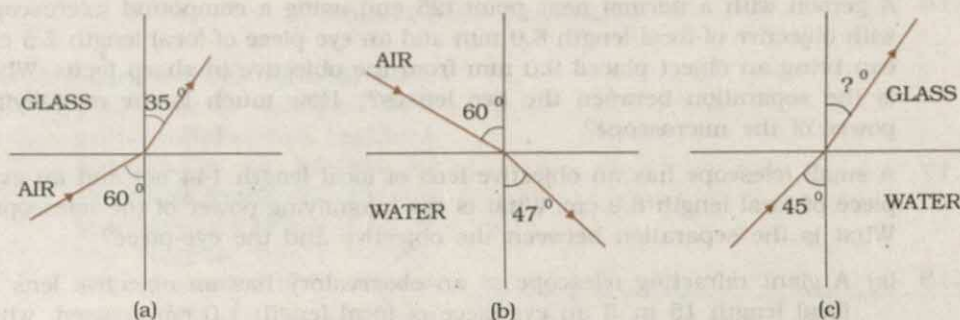


Fig. 10.33 (a) - (c)

- 10.7** A small bulb is placed at the bottom of tank containing water to a depth of 80 cm. What is the area of the surface of water through which light from the bulb can emerge out? Refractive index of water is 1.33. (Consider the bulb to be a point source.)
- 10.8** A prism is made of glass of unknown refractive index. A parallel beam of light is incident on a face of the prism. By rotating the prism, the angle of minimum deviation is measured to be 40° . What is the refractive index of the material of the prism? If the prism is placed in water (refractive index 1.33), predict the new angle of minimum deviation of a parallel beam of light. The refracting angle of the prism is 60° .
- 10.9** A needle placed 45 cm from a lens forms an image on a screen placed 90 cm on the other side of the lens. Identify the type of the lens and determine its focal length. What is the size of the image if the size of the needle is 5.0 cm?
- 10.10** Double-convex lenses are to be manufactured from a glass of refractive index 1.55, with both faces of the same radius of curvature. What is the radius of curvature required if the focal length is to be 20 cm?
- 10.11** A beam of light converges to a point P. A lens is placed in the path of the convergent beams 12 cm from P. At what point does the beam converge if the lens is (a) a convex lens of focal length 20 cm, (b) a concave lens of focal length 16 cm?
- 10.12** An object of size 3.0 cm is placed 14 cm in front of a concave lens of focal length 21 cm. Describe the image produced by the lens. What happens if the object is moved farther from the lens?
- 10.13** What is the focal length of a convex lens of focal length 30 cm in contact with a concave lens of focal length 20 cm? Is the system a converging or a diverging lens? Ignore thickness of the lenses.
- 10.14** A thin convex lens of focal length 5 cm is used as a simple microscope by a person with normal near point (25 cm). What is the magnifying power of the microscope?
- 10.15** A compound microscope consists of an objective lens of focal length 2.0 cm and an eye-piece of focal length 6.25 cm separated by a distance of 15 cm. How far from the objective should an object be placed in order to obtain the final image at (a) the least distance of distinct vision (25 cm), (b) infinity? What is the magnifying power of the microscope in each case?

- 10.16** A person with a normal near point (25 cm) using a compound microscope with objective of focal length 8.0 mm and an eye-piece of focal length 2.5 cm can bring an object placed 9.0 mm from the objective in sharp focus. What is the separation between the two lenses? How much is the magnifying power of the microscope?
- 10.17** A small telescope has an objective lens of focal length 144 cm and an eye-piece of focal length 6.0 cm. What is the magnifying power of the telescope? What is the separation between the objective and the eye-piece?
- 10.18** (a) A giant refracting telescope at an observatory has an objective lens of focal length 15 m. If an eye-piece of focal length 1.0 cm is used, what is the angular magnification of the telescope?
 (b) If this telescope is used to view the moon, what is the diameter of the image of the moon formed by the objective lens? The diameter of the moon is 3.48×10^6 m, and the radius of lunar orbit is 3.8×10^8 m.
- 10.19** A telescope has an objective of diameter 60 cm. The focal lengths of the objective and eye-piece are 2.0 m and 1.0 cm, respectively. The telescope is directed to view two distant, almost point sources of light (e.g., two stars of a binary). The sources are at roughly the same distance ($\approx 10^4$ light years) along the line of sight but are separated transverse to the line of sight by a distance of 10^{10} m. Will the telescope resolve the two objects i.e., will it see two distinct stars?
- 10.20** (a) An object is placed between two plane mirrors inclined at 60° to each other. How many images do you expect to see?
 (b) An object is placed between two plane parallel mirrors. Why do the distant images get fainter and fainter?
 (c) Why are mirrors used in search-lights parabolic and not concave spherical?
 (d) If you were driving a car, what type of mirror would you prefer to use for observing traffic at your back?
- 10.21** (a) A concave mirror and a convex lens are held in water. What change, if any, do you expect to find in the focal length of either?
 (b) On a hot summer day in a desert, one sees the reflected image of distant parts of the sky. (This is sometimes mistaken by the observer to be the reflection of the sky in some distant lake of water. This illusion is called a mirage). Explain.
 (c) What is the twinkling effect of starlight due to?
 (d) Watching the sunset on a beach, one can see the Sun for several minutes after it has 'actually set', Explain.
- 10.22** (a) People usually prefer light-coloured dresses during summer and dark dresses during winter. Why?
 (b) How would a blue object appear under sodium lamp light?
 (c) What does a welder protect against when he wears a mask?
 (d) Explain why the sky is blue, and the Sun appears red at sunset.
- 10.23** (a) What is the adjustment needed in a camera to take pictures of objects at different distances?
 (b) What does the (adjustable) f -number of a camera signify? What does one mean by saying that the aperture is $f/11$. Why are apertures labelled as $f/2$, $f/2.8$, $f/4$, $f/5.6$, $f/8$, $f/11$ etc?
 (c) What is the function of the camera shutter?

ADDITIONAL EXERCISES

- 10.24** Light incident normally on a plane mirror attached to a galvanometer coil retraces backwards as shown in Fig. 10.34. A current in the coil produces a deflection of 3.5° of the mirror. What is the displacement of the reflected spot of light on a screen placed 1.5 m away?

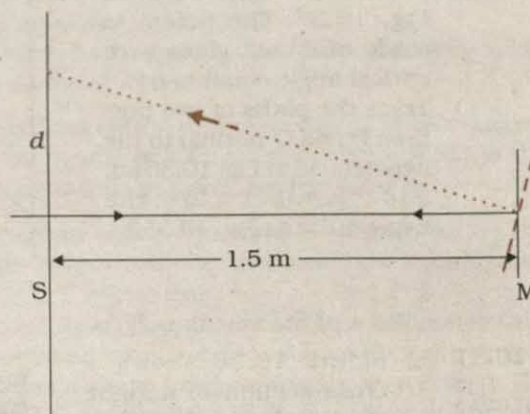


Fig. 10.34

- 10.25** A boy 1.50 m tall with his eye-level at 1.38 m stands before a mirror fixed on a wall. Indicate by means of a ray diagram how the mirror should be positioned so that he can view himself fully. What should be the minimum length of the mirror? Does the answer depend on the eye-level?
- 10.26** Light incident on a rotating mirror M is reflected to a fixed mirror N placed 22.5 km away from M. The fixed mirror reflects it back to M (along the same path) which in turn reflects the light again along a direction that makes an angle of 27° with the incident direction. What is the speed of rotation of the mirror if the speed of light is $3.0 \times 10^8 \text{ m s}^{-1}$? (Principle of Foucault's method for the determination of c).
- 10.27** Use the mirror equation to deduce that:
- an object placed between f and $2f$ of a concave mirror produces a real image beyond $2f$.
 - a convex mirror always produces a virtual image independent of the location of the object.
 - the virtual image produced by a convex mirror is always diminished in size and is located between the focus and the pole.
 - an object placed between the pole and focus of a concave mirror produces a virtual and enlarged image.
- [Note: This exercise helps you deduce algebraically properties of images that one obtains from explicit ray diagrams.]
- 10.28** A small pin fixed on a table top is viewed from above from a distance of 50 cm. By what distance would the pin appear to be raised if it is viewed from the same point through a 15 cm thick glass slab held parallel to the table? Refractive index of glass = 1.5. Does the answer depend on the location of the slab?
- 10.29** The bottom of a container is a 4.0 cm thick glass ($n = 1.5$) slab. The container contains two immiscible liquids A and B of depths 6.0 cm and 8.0 cm, respectively. What is the apparent position of a scratch on the outer surface of the bottom of the glass slab when viewed through the container? Refractive indices of A and B are 1.4 and 1.3, respectively.

- 10.30** A right-angle prism is placed before an object in the two positions shown in Fig. 10.35. The prism is made of crown glass with critical angle equal to 41° . Trace the paths of two rays from P and Q normal to the hypotenuse in Fig. 10.35 (a), and parallel to the hypotenuse in Fig. 10.35(b).

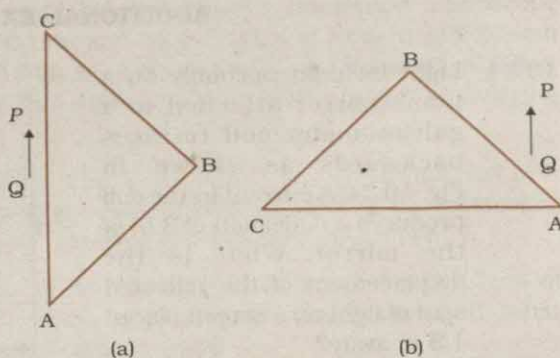


Fig. 10.35

- 10.31** (a) Figure 10.36 shows a cross-section of a 'light pipe' made of a glass fibre of refractive index 1.68. The outer covering of the pipe is made of a material of refractive index 1.44. What is the range of the angles of the incident rays with the axis of the pipe for which total reflections inside the pipe take place as shown in the figure.

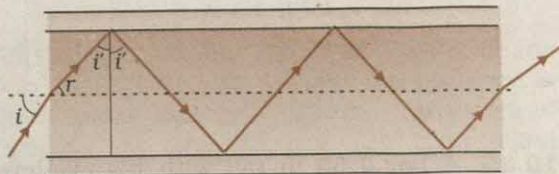


Fig. 10.36

- (b) What is the answer if there is no outer covering of the pipe?
- 10.32** Parallel light from the collimator of a spectrometer is incident on the two faces of a prism which make the refracting angle A of the prism. The image of the collimator slit is observed in two different positions of the telescope of the spectrometer. If the angle of rotation of the telescope between the two positions is 144° , what is the angle A of the prism?
- 10.33** Use the lens equation to deduce algebraically what you know otherwise from explicit ray diagrams:
- An object placed within the focus of a convex lens produces a virtual and enlarged image.
 - A concave lens produces a virtual and diminished image independent of the location of the object.
- 10.34** Answer the following questions:
- A man holding a lighted candle in front of a thick glass mirror and viewing it obliquely sees a number of images of the candle. What is the origin of these multiple images?
 - You read a newspaper because of the light that it reflects. Then why do you not see even a faint image of yourself in the newspaper?
 - You have learnt that plane and convex mirrors produce virtual images of objects. Can they produce real images under some circumstances? Explain.
 - The wall of a room is covered with a perfect plane mirror, and two movie films are made, one recording the movement of a man and the other of his mirror image. From viewing the films later, can an outsider tell which is which?
- 10.35** Answer the following questions:
- A virtual image, we always say, cannot be caught on a screen. Yet when we 'see' a virtual image, we are obviously bringing it on to the 'screen' (i.e., the retina) of our eye. Is there a contradiction?

- (b) To a fish under water, viewing obliquely a fisherman standing on the bank of a lake, does the man look taller or shorter than what he actually is?
- (c) Does the apparent depth of a tank of water change if viewed obliquely? If so, does the apparent depth increase or decrease?
- (d) The refractive index of diamond is much greater than that of ordinary glass. Is this fact of some use to a diamond cutter?

10.36 The image of a small electric bulb fixed on the wall of a room is to be obtained on the opposite wall 3 m away by means of a large convex lens. What is the maximum possible focal length of the lens required for the purpose?

10.37 (a) A screen is placed 90 cm from an object. The image of the object on the screen is formed by a convex lens at two different locations separated by 20 cm. Determine the focal length of the lens.

- (b) Suppose the object in (a) above, is an illuminated slit in a collimator tube so that it is hard to measure the slit size and its distance from the screen. Using a convex lens one obtains a sharp image of the slit on a screen. The image size is measured to be 4.6 cm. The lens is displaced away from the slit and at a certain location, another sharp image of size 1.7 cm is obtained. Determine the size of the slit.

10.38 (a) Determine the 'effective focal length' of the combination of the two lenses in Exercise 10.13 if they are placed 8.0 cm apart with their principal axes coincident. Does the answer depend on which side a beam of parallel light is incident? Is the notion of effective focal length of this system useful at all?

- (b) An object 1.5 cm in size is placed on the side of the convex lens in the above arrangements. The distance between the object and the convex lens is 40 cm. Determine the magnification produced by the two-lens system, and the size of the image.

10.39 At what angle should a ray of light be incident on the face of a prism of refracting angle 60° so that it just suffers total internal reflection at the other face? The refractive index of the material of the prism is 1.524.

10.40 Figure 10.37 shows an equiconvex lens (of refractive index 1.50) in contact with a liquid layer on top of a plane mirror. A small needle with its tip on the principal axis is moved along the axis until its inverted image is found at the position of the needle. The distance of the needle from the lens is measured to be 45.0 cm. The liquid is removed and the experiment is repeated. The new distance is measured to be 30.0 cm. What is the refractive index of the liquid?

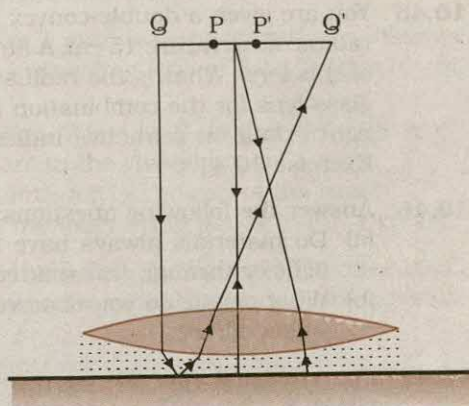


Fig. 10.37

10.41 Using a spectrometer, the following data are obtained for a crown glass prism and a flint glass prism.

Crown glass prism:

Angle of the Prism, $A = 72.0^\circ$

Angle of minimum deviation:

$$\delta_b = 54.6^\circ,$$

$$\delta_r = 53.0^\circ,$$

$$\delta_y = 54.0^\circ,$$

Flint glass prism:

$$A = 60.0^\circ$$

$$\delta_b = 52.8^\circ,$$

$$\delta_r = 50.6^\circ,$$

$$\delta_y = 51.9^\circ,$$

where b, r and y refer to particular wavelengths in the blue, red and yellow bands. Compare the dispersive powers of the two varieties of glass prisms.

- 10.42** You are given prisms made of crown glass and flint glass with a wide variety of angles. Suggest a combination of prisms which will (a) deviate a pencil of white light without much dispersion, (b) disperse (and displace) a pencil of white light without much deviation. Use qualitatively the answer to Exercise 10.41.
- 10.43** A convex lens made of a variety of glass of high dispersive power has focal length of 15 cm. A parallel beam of white light is incident on one side of the lens and screen is placed on the other side. Describe the chromatic aberration of the lens, i.e., describe the colours on the spot focussed on the screen as the screen is moved away from the lens.
- 10.44** (a) A combination of two thin lenses in contact is to be made which has the same focal length for blue and red light. (Such a combination is known as an 'achromatic doublet'.) Show that the ratio of their focal lengths (for yellow light) must be equal in magnitude and opposite in sign to the ratio of the dispersive powers of the materials of the two lenses.
 (b) Use the results in (a) to suggest a way of removing chromatic aberration of the lens in Exercise 10.43 which is made of flint glass. You are given convex and concave lenses (made of crown glass) of various focal lengths. The ratio of the dispersive powers of flint glass to crown glass is about 1.5.
- 10.45** You are given a double-convex lens made of crown glass with each surface of radius of curvature 15 cm. A flint glass lens is grafted on to one of the surfaces of this lens. What is the radius of curvature of the second surface of the flint glass lens for the combination to be an 'achromatic doublet' for blue and red light? Data on refractive indices required may be obtained from answer to Exercise 10.41.
- 10.46** Answer the following questions:
 (a) Do materials always have the same colour whether viewed by reflected light or through transmitted light?
 (b) What colour do you observe when white light passes through a blue and yellow filter?
- 10.47** For a normal eye, the far point is at infinity and the near point of distinct vision is about 25 cm in front of the eye. The cornea of the eye provides a converging power of about 40 dioptres, and the least converging power of the eye-lens behind the cornea is about 20 dioptres. From this rough data estimate the range of accommodation (i.e., the range of converging power of the eye-lens) of a normal eye.
- 10.48** Does short-sightedness (myopia) or long-sightedness (hypermetropia) imply necessarily that the eye has partially lost its ability of accommodation? If not, what might cause these defects of vision?

- 10.49** (a) The far point of a myopic person is 80 cm in front of the eye. What is the power of the lens required to enable him to see very distant objects clearly?
 (b) In what way does the corrective lens help the person above? Does the lens magnify very distant objects? Explain carefully.
 (c) The person above prefers to remove his spectacles while reading a book. Explain why?
- 10.50** (a) The near point of a hypermetropic person is 75 cm from the eye. What is the power of the lens required to enable him to read clearly a book held at 25 cm from the eye?
 (b) In what way does the corrective lens help the person above? Does the lens magnify objects held near the eye?
 (c) The person above prefers to remove his spectacles while looking at the sky. Explain why?
- 10.51** (a) Suppose the person in Exercise 10.49 uses spectacles of power -0.80 dioptre, how far can he see clearly?
 (b) If the person in Exercise 10.50 uses spectacles of power $+1.0$ dioptre, what is the nearest distance of distinct vision for him?
- 10.52** A myopic person has been using spectacles of power -1.0 dioptre for distant vision. During old age he also needs to use separate reading glass of power $+2.0$ dioptres. Explain what may have happened.
- 10.53** A person looking at a mesh of crossed wires is able to see the vertical wires more distinctly than the horizontal wires. What is this defect due to? How is such a defect of vision corrected?
- 10.54** A man with normal near point (25 cm) reads a book with small print using a magnifying glass: a thin convex lens of focal length 5 cm.
 (a) What is the closest and the farthest distance at which he can read the book when viewing through the magnifying glass?
 (b) What is the maximum and the minimum angular magnification (magnifying power) possible using the above simple microscope?
- 10.55** A figure divided into squares each of size 1 mm^2 is being viewed at a distance of 9 cm through a magnifying glass (a converging lens of focal length 10 cm) held close to the eye.
 (a) What is the magnification (image size/object size) produced by the lens? How much is the area of each square in the virtual image?
 (b) What is the angular magnification (magnifying power) of the lens?
 (c) Is the magnification in (a) equal to the magnifying power in (b)? Explain.
- 10.56** (a) At what distance should the lens be held from the figure in Exercise 10.55 in order to view the squares distinctly with the maximum possible magnifying power?
 (b) What is the magnification (image size/object size) in this case?
 (c) Is the magnification equal to the magnifying power in this case? Explain.
- 10.57** What should be the distance between the object in Exercise 10.55 and the magnifying glass if the virtual image of each square in the figure is to have an area of 6.25 mm^2 . Would you be able to see the squares distinctly with your eyes very close to the magnifier?
 [Note: Exercises 10.55 to 10.57 will help you clearly understand the difference between magnification in absolute size and the angular magnification (or magnifying power) of an instrument.]

10.58 Answer the following questions:

- The angle subtended at the eye by an object is equal to the angle subtended at the eye by the virtual image produced by a magnifying glass. In what sense then does a magnifying glass provide angular magnification?
- In viewing through a magnifying glass, one usually positions one's eyes very close to the lens. Does angular magnification change if the eye is moved back?
- Magnifying power of a simple microscope is inversely proportional to the focal length of the lens. What then stops us from using a convex lens of smaller and smaller focal length and achieving greater and greater magnifying power?
- Why must both the objective and the eye-piece of a compound microscope have short focal lengths?
- When viewing through a compound microscope, our eyes should be positioned not on the eye-piece but a short distance away from it for best viewing. Why? How much should be that short distance between the eye and eye-piece?

10.59 An angular magnification (magnifying power) of 30X is desired using an objective of focal length 1.25 cm and an eye-piece of focal length 5 cm. How will you set up the compound microscope?

10.60 A small telescope has an objective lens of focal length 140 cm and an eye-piece of focal length 5.0 cm. What is the magnifying power of the telescope for viewing distant objects when

- the telescope is in normal adjustment (i.e., when the final image is at infinity)?
- the final image is formed at the least distance of distinct vision (25 cm)?

10.61 (a) For the telescope described in Exercise 10.60 (a), what is the separation between the objective lens and the eye-piece?
 (b) If this telescope is used to view a 100 m tall tower 3 km away, what is the height of the image of the tower formed by the objective lens?
 (c) What is the height of the final image of the tower if it is formed at 25 cm?

10.62 An amateur astronomer wishes to estimate roughly the size of the Sun using his crude telescope consisting of an objective lens of focal length 200 cm and an eye-piece of focal length 10 cm. By adjusting the distance of the eye-piece from the objective, he obtains an image of the Sun on a screen 40 cm behind the eye-piece. The diameter of the Sun's image is measured to be 6.0 cm. What is his estimate of the Sun's size, given that the average Earth-Sun distance is 1.5×10^{11} m?

10.63 (a) List some advantages of a reflecting telescope, especially for high resolution astronomy.
 (b) A reflecting telescope has a large mirror for its objective with radius of curvature equal to 80 cm. What is the magnifying power of the telescope if the eye-piece used has a focal length of 1.6 cm?

10.64 (a) The image of the objective in the eye-piece is known as the 'eye-ring'. Why is this the best position of our eyes for viewing?
 (b) Show that the angular magnification of a telescope equals the ratio of the diameter of objective to the diameter of eye-ring.
 (c) The angular magnification of a telescope is 300. What should be the diameter of the objective if our eyes (located at the eye-ring) are just able to collect all the light refracted by the objective? Take the diameter of the pupil of eye to be 3 mm.

- 10.65** A terrestrial telescope has an objective of focal length 180 cm and an eye-piece of focal length of 5 cm. The erecting lens has a focal length of 3.5 cm. What is the separation between the objective and the eye-piece? What is the magnifying power of the telescope? Can we use the telescope for viewing astronomical objects?
- 10.66** (a) A Galilean telescope obtains the final image erect (like in a terrestrial telescope) without an intermediate erecting lens. It does so by using a diverging lens for its eye-piece. Show that the angular magnification of a Galilean telescope is given by a formula similar to that for any ordinary telescope: angular magnification $= -f_o/f_e$ (negative sign because f_e is negative).
 (b) For a Galilean telescope with $f_o = 150$ cm, $f_e = -7.5$ cm, what is the separation between the objective and the eye-piece?
 (c) What is the main disadvantage of this type of telescope?
- 10.67** Describe briefly the construction of prism binoculars. Explain with the help of a ray diagram how the inversion of the image produced by an ordinary telescope is reversed by the use of two right-angled totally reflecting prisms. Explain carefully how the final image is both erect and without any 'lateral inversion'. List some advantages of prism binoculars over an ordinary telescope.
- 10.68** The objective of telescope A has a diameter 3 times that of the objective of telescope B.
 (a) How much greater amount of light is gathered by A compared to B?
 (b) Show that the range of A is three times the range of B. (Range tells you how far away a star of some standard absolute brightness can be spotted by the telescope.)
 (c) A telescope increases the brightness of the background compared to what is seen by the unaided eye. Thus, it facilitates observation by improving the contrast between a star and its background. Explain this statement carefully.
- 10.69** A 35 mm slide with a 24 mm \times 36 mm picture is projected on a screen placed 12 m from the slide. The image of the slide picture on the screen measures 1.0 m \times 1.5 m. Determine the location of the projection lens, and its focal length.
- 10.70** In order to help you understand clearly the functions of the condensing lens and the projection lens in a projector, the following questions have been devised. Answer them carefully:
 (a) Why do we need a condensing lens at all? Can we not project a slide simply by illuminating it by a lamp and obtaining its magnified image on a screen?
 (b) For projecting the slide in Exercise 10.69, how much should be the least diameter of the condensing lens? Where should the slide be placed relative to this lens?
 (c) The image of the source formed by the condensing lens should be located on the projection lens. Why? What is the preferred size of this image of the source? Explain.
 (d) The condensing lens converges light from the source on to the slide. Does that mean the image of the source is formed on the slide? If not, why not?
 (e) Consider a small portion of the image on the screen, say the lowermost portion. Does this portion get illumination only from the lowermost part of the source? Or from all points of the source? Explain.
- 10.71** The following questions will help you clearly understand the simple optical principles involved in a camera. Answer them carefully:
 (a) Explain the term 'depth of field'. Why does the depth of field increase if aperture is reduced? Which shot in your view will require a greater depth of view—photograph of a scenic spot, or your identity photograph?
 (b) The field of view of a camera is increased by using a so called 'wide-angle

lens'. In what way does this lens differ from an ordinary camera lens? How does it increase the field of view?

- (c) What is a telephoto lens? How does it differ from an ordinary camera lens? How does it increase the field of view?
- (d) Why are apertures of camera lenses so small while the apertures of telescopes are as large as feasible?

10.72 (a) Show that for a given brightness of the image on a camera film, the exposure time t is inversely proportional to the square of the aperture size a and directly proportional to the square of the focal length f of the camera lens.

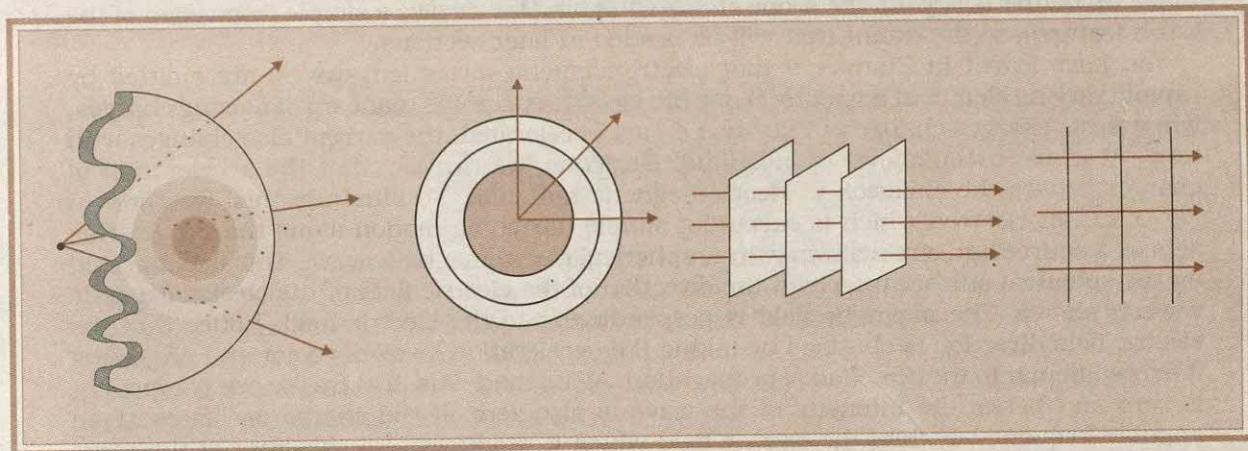
- (b) A camera is set at the aperture size $f/8$ and the exposure time of $(1/60)\text{s}$. How much exposure time is required for receiving the same amount of light if the aperture size is set at $f/5.6$? How is the depth of field affected by this change?

10.73 An eye-piece of a telescope consists of two plano-convex lenses L_1 and L_2 each of focal length f separated by a distance of $2f/3$. Where should L_1 be placed relative to the focus of the objective lens of the telescope so that the final image through L_2 is seen at infinity?

[Note: An arrangement of lenses like this is preferred to a simple double-convex lens for an eye-piece because it reduces chromatic and spherical aberrations. Details are beyond our scope here.]

CHAPTER ELEVEN

WAVE OPTICS



11.1 INTRODUCTION

Starting from the seventeenth century, new kinds of behaviour of light were discovered experimentally which could not be understood using simple ray concepts. These experiments were explained in terms of light *waves*. Later, in the nineteenth century, the study of electricity and magnetism led to Maxwell's idea of electromagnetic waves. These were experimentally verified by Hertz (Chapter 9). It was realised by Maxwell and Hertz that visible light is also made up of waves of this kind. The wavelength varies from approximately 390 nm (3.9×10^{-7} m) for violet to 760 nm (7.6×10^{-7} m) for red.

The wave-like behaviour of light is the subject of this chapter. Some of the effects described are common to all kinds of waves including sound waves. We all know that we can hear a person speaking even when we cannot see him. This means that the energy of sound waves is able to travel even when there is no straight line path from the source to the observer. The same phenomenon occurs when light falls on a small slit or hole. The energy of the light wave spreads out. This phenomenon is called *diffraction* of light. We do not notice diffraction of light in our daily life because the wavelength of light is so short.

Another remarkable effect is *interference*, which you have already come across (Chapter 15, Class XI). Waves from two different sources reaching the same point can combine to produce an intensity which is not the sum of the individual intensities. In fact, the result can even be zero intensity. Superposition of two light waves can produce darkness. This is called *destructive interference*. The resultant intensity can also be more than the sum of the individual intensities. Such a situation is called *constructive interference*.

The *Doppler effect* is again common to all waves. When there is relative motion between the source and the observer, the frequency observed is different from the frequency emitted. You have already encountered this in your study of sound waves (Chapter 15, Class XI).

There are some interesting effects which occur only for transverse waves. As we have seen in Chapter 9, the electric field of light waves is perpendicular to the direction of propagation. In other words, the electric field vector lies in same direction in the transverse plane. It is found experimentally that the behaviour of light in reflection, refraction, and transmission through crystals depends on the direction of the electric field in the plane perpendicular to the direction of travel. This property is called *polarisation* of light. Polarisation is the subject of the last section of this Chapter.

11.1.1 Emission, Absorption, and Scattering of Light Waves

A detailed description of how light waves interact with the electrons in the atoms which make up matter is beyond the scope of this Chapter. But we give a simple discussion of the basic features, to the extent that will be needed in later sections.

You have learnt in Chapter 9 that electromagnetic waves (em waves) are emitted by rapidly varying electrical currents. Since the electric current is made up of moving charges, when these charges change in velocity, i.e., are accelerated, the current also changes with time. At a microscopic level of individual electrons, we can say that the acceleration of charges causes the emission of electromagnetic radiation. To illustrate this, we show in Fig. 11.1, an electron which is executing simple harmonic motion along the z -axis. This acts as a source and emits a transverse, spherical electromagnetic wave. This is illustrated by the spherical surface on which the direction of the electric field of the electromagnetic wave is shown. The magnetic field is perpendicular to the electric field. Notice that the electric field direction is obtained by taking the acceleration locus of vector and projecting it perpendicular to the direction of propagation. Along the z -axis, this transverse component is zero and hence the intensity of the wave is also zero. If the charge oscillates at an angular frequency ω , then the electric field which it produces also oscillates at the same frequency, and hence this is the frequency of the wave.

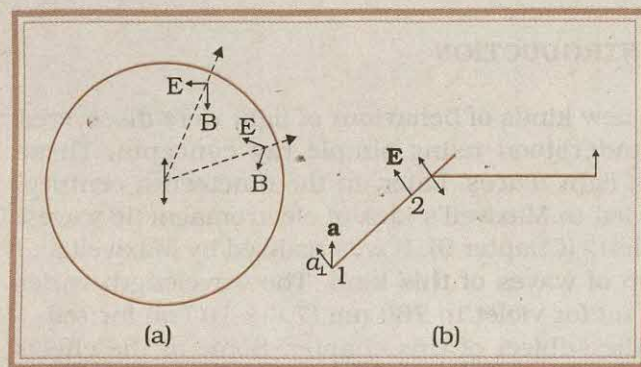


Fig. 11.1 (a) Spherical wave emitted by an electron oscillating along the z -axis. Note the \mathbf{E} and \mathbf{B} vectors. (b) Emission, absorption and scattering of em waves.

This simple model of emission of electromagnetic waves will be enough for our purposes in this Chapter. We can think of it as transfer of energy from the accelerated electron to the em wave. The reverse kind of energy transfer is also possible. When such a wave falls on another electron, which is initially at rest, then the electron is set into motion because of the force which is exerted on it by the electric field of the wave. (The magnetic field also exerts a force on a moving electron, but for charges which move much slower than light, the magnetic force exerted by the em wave is much smaller than the

electric force). The energy taken up by the second electron which is set into motion is removed from the incident wave. This process is called *absorption* of radiation.

Further, when the second electron is accelerated, it too will emit radiation. In this process, another spherical em wave, with the second electron as centre, is produced. This is called *scattering* of em radiation. As the figure shows, the energy from the first electron can reach the observer via the second electron. In this process, the direction of travel of the energy changes. This process is therefore called *scattering* of radiation.



Christiaan Huygens **(1629-1695)**

Dutch physicist, astronomer, mathematician and the founder of the wave theory of light. His book, "Treatise on light", makes fascinating reading even today. He brilliantly explained the double refraction shown by the mineral calcite in this work in addition to reflection and refraction. He was the first to analyse circular and simple harmonic motion and designed and built improved clocks and telescopes. He discovered the true geometry of Saturn's rings.



Thomas Young **(1773-1829)**

English physicist, physician and Egyptologist. Young worked on a wide variety of scientific problems, ranging from the structure of the eye and the mechanism of vision to the decipherment of the Rosetta stone. He revived the wave theory of light and recognized that interference phenomena provide proof of the wave properties of light.

11.2 WAVEFRONTS, RAYS, AND HUYGENS' PRINCIPLE

11.2.1 Wavefronts

Consider a wave spreading out on the surface of water after a stone is thrown in. Every point on the surface oscillates. At any time, a photograph of the surface would show circular rings on which the disturbance is maximum. Clearly, all points on such a circle are oscillating in phase because they are at the same distance from the source. Such a locus of points which oscillate in phase is an example of a *wavefront*. **A wavefront is defined as a surface of constant phase.** The speed with which the wavefront moves outwards from the source is called the *phase speed*. The energy of the wave travels in a direction perpendicular to the wavefront.

Figure 11.2 shows light waves from a point source forming a spherical wavefront in three dimensional space. The energy travels outwards along straight lines emerging from the source, i.e., radii of the spherical wavefront. These lines are the rays. Notice that when we measure the spacing between a pair of wavefronts along any ray, the result is a constant. This example illustrates two important general principles which we will use later:

- (i) *Rays are perpendicular to wavefronts,*
- (ii) *The time taken for light to travel from one wavefront to another is the same along any ray.*

If we look at a small portion of a spherical wave, far away from the source, then the wavefronts are like parallel planes. The rays are

parallel lines perpendicular to the wavefronts. This is called a plane wave and is also sketched in Fig. 11.2.

A linear source such as a slit illuminated by another source behind it will give rise to cylindrical wavefronts. Again, at larger distance from the source, these 'wave fronts may be regarded as planar'.

11.2.2 Huygens' Construction

Huygens, Dutch physicist and astronomer of the seventeenth century, gave a beautiful geometrical description of wave propagation. We can guess that he must have seen water waves many times in the canals of his native place Holland. A stick placed in water and oscillated up and down becomes a source of waves. Since the surface of water is two dimensional, the resulting wavefronts would be circles instead of spheres. At each point on such a circle, the water level moves up and down. Huygens' idea is that we can think of every such oscillating point on a wavefront as a new source of waves. According to Huygens' principle, what we observe is the result of adding up the waves from all these different sources. These are called secondary waves or wavelets.

Huygens' principle is illustrated in Fig. 11.3, in the simple case of a plane wave.

1. At time $t = 0$, we have a wavefront F_1 . F_1 separates those parts of the medium which are undisturbed from those where the wave has already reached.
2. Each point on F_1 acts like a new source and sends out a spherical wave. After a time t ,

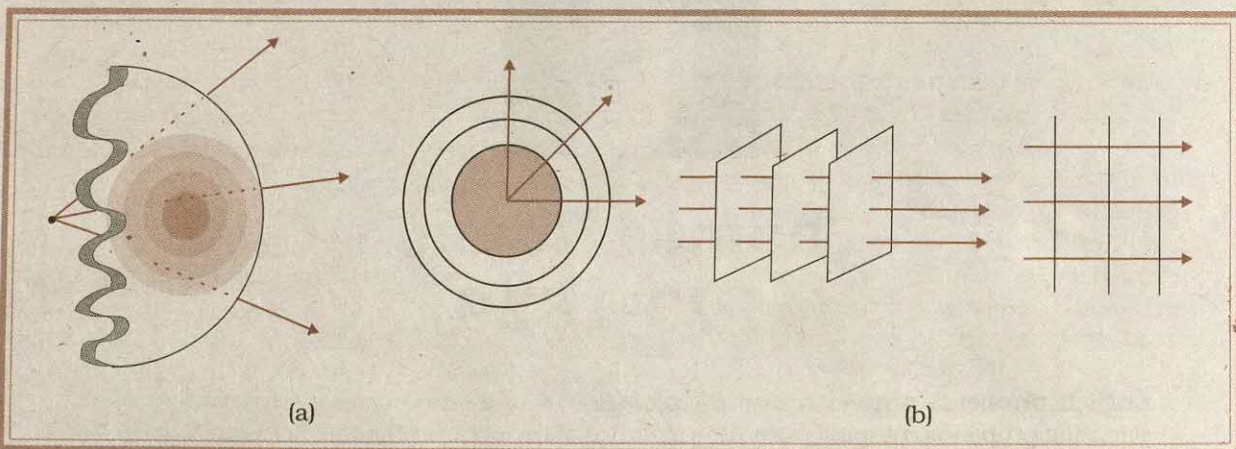


Fig. 11.2 Wavefronts and the corresponding rays in two cases: (a) diverging spherical wave, (b) plane wave. The figure on the left shows a wave (e.g., light) in three dimensions. The figure on the right shows a wave in two dimensions (a water surface).

- each of these will have radius vt . These spheres are the secondary wavelets.
- After a time t , the disturbance would now have reached all points within the region covered by all these secondary waves. The boundary of this region is the new wavefront F_2 . Notice that F_2 is a surface tangent to all the spheres. It is called the *forward envelope* of these secondary wavelets.
 - The secondary wavelet from the point A_1 on F_1 touches F_2 at A_2 . Draw the line connecting any point A_1 on F_1 to the corresponding point A_2 on F_2 . According to Huygens, A_1A_2 is a ray. It is perpendicular to the wavefronts F_1 and F_2 and has length vt . This implies that rays are perpendicular to wavefronts. Further, the time taken for light to travel between two wavefronts is the same along any ray. In our example, the speed v of the wave has been taken to be the same at all points in the medium. In this case, we can say that the distance between two wavefronts is the same, measured along any ray.
 - This geometrical construction can be repeated starting with F_2 to get the next wavefront F_3 a time t later, and so on. It is known as Huygens' construction.

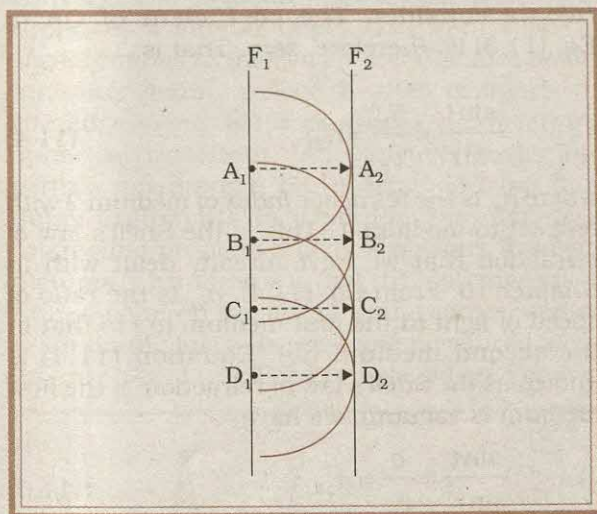


Fig. 11.3 Huygens' geometrical construction for wave propagation. F_1 is a wavefront at some time, and F_2 at a time t later. The lines A_1A_2 , B_1B_2 etc., are normal to both F_1 and F_2 and represent rays.

Huygens' construction can be understood physically for waves in a material medium, like the surface of water. Each oscillating particle can

set its neighbours into oscillation, and therefore acts as a secondary source. But what if there is no medium, as for light travelling in vacuum? The mathematical theory, which cannot be given here, shows that the same geometrical construction works in this case as well.

11.2.3 Reflection and Refraction

We can use a modified form of Huygens' construction to understand reflection and refraction of light. Figure 11.4(a) shows an incident wavefront which makes an angle i with the surface separating two media, for example, air and water. The phase speeds in the two media are v_1 and v_2 . We can see that when the point A on the incident wavefront strikes the surface, the point B still has to travel a distance $BC = AC \sin i$, and this takes a time $t = BC/v_1 = AC (\sin i)/v_1$. After a time t , a secondary wavefront of radius v_2t with A as centre would have travelled into medium 2. The secondary wavefront with C as centre would have just started, i.e., would have zero radius. We also show a secondary wavelet originating from a point D in between A and C. Its radius is less than v_2t . The wavefront in medium 2 is thus a line passing through C and tangent to the circle centred on A. We can see that the angle r' made by this refracted wavefront with the surface is given by $AE = v_2t = AC \sin r'$. Hence, $t = AC (\sin r')/v_2$. Equating the two expressions for t gives us the law of refraction in the form $\sin i / \sin r' = v_1 / v_2$. A similar picture is drawn in Fig. 11.4(b) for the reflected wave which travels back into medium 1. In this case, we denote the angle made by the reflected wavefront with the surface by r , and we find that $i = r$. Notice that for both reflection and refraction, we use secondary wavelets starting at different times. Compare this with the earlier application (Fig. 11.3) where we start them at the same time.

The preceding argument gives a good physical picture of how the refracted and reflected waves are built up from secondary wavelets. We can also understand the laws of reflection and refraction using the concept that the time taken by light to travel along different rays from one wavefront to another must be the same. Figure 11.4(c) shows the incident and reflected wavefronts when a parallel beam of light falls on a plane surface. One ray PQ is shown

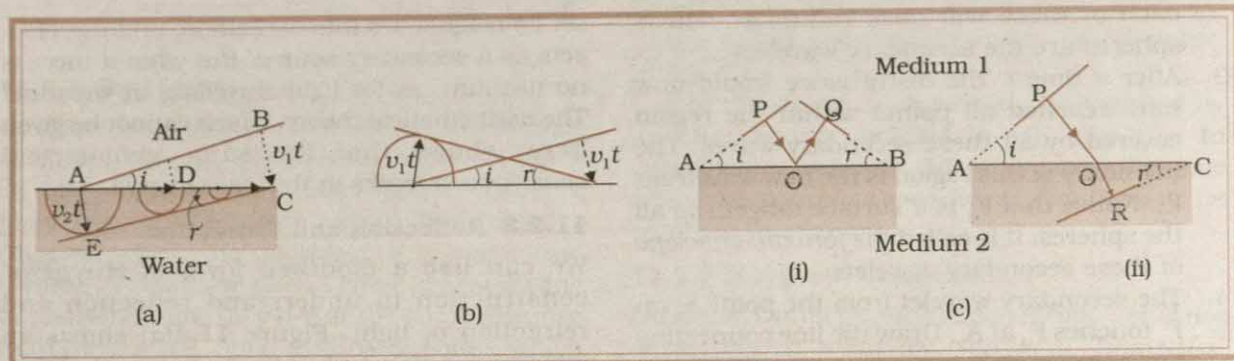


Fig. 11.4 (a) Huygens' construction for the (a) refracted wave. (b) Reflected wave. (c) Calculation of propagation time between wavefronts in (i) reflection and (ii) refraction.

normal to both the reflected and incident wavefronts. The angle of incidence i and the angle of reflection r are defined as the angles made by the incident and reflected rays with the normal. As shown in Fig. 11.4(c), these are also the angles between the wavefront and the surface.

We now calculate the total time to go from one wavefront to another along the rays. From Fig. 11.4(c), we have,

Total time for light to reach from P to Q

$$\begin{aligned} &= \frac{PO}{v_1} + \frac{OQ}{v_1} = \frac{AO \sin i}{v_1} + \frac{OB \sin r}{v_1} \\ &= \frac{OA \sin i + (AB - OA) \sin r}{v_1} \\ &= \frac{AB \sin r + OA (\sin i - \sin r)}{v_1} \quad (11.1) \end{aligned}$$

Different rays normal to the incident wavefront strike the surface at different points O and hence have different values of OA. Since the time should be the same for all the rays, the right side of Eq. (11.1) must actually be independent of OA. The condition for this to happen is that the coefficient of OA in Eq. (11.1) should be zero, i.e., $\sin i = \sin r$. We, thus, have the law of reflection,

$$i = r \quad (11.2)$$

Figure 11.4 also shows refraction at a plane surface separating medium 1 (speed of light v_1) from medium 2 (speed of light v_2). The incident and refracted wavefronts are shown, making angles i and r' with the boundary. Angle r' is called the angle of refraction. Rays perpendicular to these are also drawn. As before, let us

calculate the time taken to travel between the two wavefronts along any ray.

$$\begin{aligned} \text{Time taken from P to R} &= \frac{PO}{v_1} + \frac{OR}{v_2} \\ &= \frac{OA \sin i}{v_1} + \frac{(AC - OA) \sin r'}{v_2} \\ &= \frac{AC \sin r'}{v_2} + OA \left(\frac{\sin i}{v_1} - \frac{\sin r'}{v_2} \right) \quad (11.3) \end{aligned}$$

This time should again be independent of which ray we consider. The coefficient of OA in Eq. (11.3) is, therefore, zero. That is,

$$\frac{\sin i}{\sin r'} = \frac{v_1}{v_2} = n_{21} \quad (11.4)$$

where n_{21} is the *refractive index* of medium 2 with respect to medium 1. This is the Snell's law of refraction that we have already dealt with in Chapter 10. From Eq. (11.4), n_{21} is the ratio of speed of light in the first medium (v_1) to that in the second medium (v_2). Equation (11.4) is known as the Snell's law of refraction. If the first medium is vacuum, we have

$$\frac{\sin i}{\sin r'} = \frac{c}{v_2} = n_2 \quad (11.5)$$

where n_2 is the refractive index of medium 2 with respect to vacuum, also called the *absolute refractive index* of the medium. A similar equation defines absolute refractive index n_1 of the first medium. From Eq. (11.4), we then get

$$n_{21} = \frac{v_1}{v_2} = \left(\frac{c}{n_1} \right) / \left(\frac{c}{n_2} \right) = \frac{n_2}{n_1} \quad (11.6)$$

The absolute refractive index of air is about 1.0003, quite close to 1. Hence, for all practical purposes, absolute refractive index of a medium may be taken with respect to air. For water,

$n_1 = 1.33$, which means $v_1 = \frac{c}{1.33}$, i.e., about

0.75 times the speed of light in vacuum. The measurement of the speed of light in water by Foucault (1850) confirmed this prediction of the wave theory.

Once we have the laws of reflection and refraction, the behaviour of prisms, lenses, and mirrors can be understood. These topics are discussed in detail in the previous Chapter. Here we just describe the behaviour of the wavefronts in these three cases (Fig. 11.5). (i) Consider a plane wave passing through a thin prism. Clearly, the portion of the incoming wavefront which travels through the greatest thickness of glass has been delayed the most, since light travels more slowly in glass. This explains the tilt in the emerging wavefront. (ii) Similarly, the central part of an incident plane wave traverses the thickest portion of a convex lens and is delayed the most. The emerging wavefront has a depression at the centre. It is spherical and converges to a focus. (iii) A concave mirror produces a similar effect. The centre of the wavefront has to travel a greater distance before and after getting reflected, when compared to the edge. This again produces a converging spherical wavefront. (iv) Concave lenses and convex mirrors can be understood from time delay arguments in a similar manner. One interesting property which is obvious from the pictures of wavefronts is that the total time taken from a point on the object to the corresponding point on the image is the same measured along any ray (Fig. 11.5). For example, when a convex

lens focuses light to form a real image, it may seem that rays going through the centre are shorter. But because of the slower speed in glass, the time taken is the same as for rays travelling near the edge of the lens.

11.2.4 Wavelength and Frequency

Let us consider a source of light at rest in one medium with the observer at rest in another medium. Further, let there be no relative motion between the two media so that the geometry of source, medium and observer does not change with time. The time taken to travel between source and observer is then fixed. Let two wavefronts separated by a whole cycle of phase (i.e., 2π) be emitted from the source, separated by a time T . Their arrival at the observer is also separated by the same time interval T . The frequency $\nu = 1/T$, therefore, remains the same as light travels from one medium to another (as frequency is the characteristic of the source) i.e., $\nu_1 = \nu_2$. Because the speeds of light v_1 and v_2 are different, the wavelengths λ_1 and λ_2 are also different. Using the relation $v = \nu\lambda$, we have

$$\frac{n_2}{n_1} = \frac{v_1}{v_2} = \frac{v_1 \lambda_1}{v_2 \lambda_2} = \frac{\lambda_1}{\lambda_2} \quad (11.7)$$

The wavelength in a medium is directly proportional to the phase speed and hence inversely proportional to the refractive index.

We should note that this reasoning will change if either the source, or the observer, is moving or there is a relative motion between the media. In that case, each successive wavefront may take a different time for its journey, when compared to the previous one. For example, if there is no medium but the source moves away from the observer, then later wavefronts have to travel a greater distance and hence take a longer time. The time period between two successive

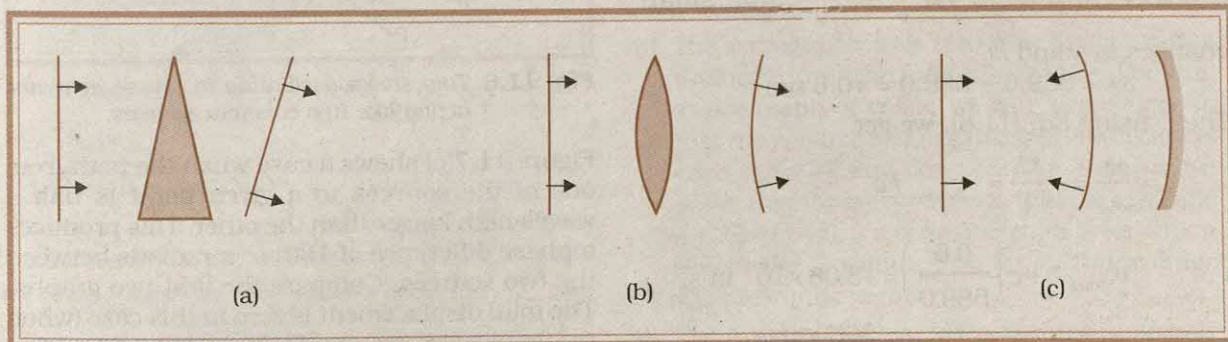


Fig. 11.5 Action on a plane wavefront due to (a) a thin prism, (b) a convex lens, and (c) a concave mirror.

wavefronts is hence longer at the observer than it is at the source. In this case, a decrease of frequency is caused by the recession of the source from the observer. This is an example of the *Doppler effect*. Astronomers call this increase in wavelength 'red shift' since a wavelength in the middle of the visible region of the spectrum moves towards the red end. A decrease in wavelength, caused by motion of the source towards the observer, is called 'blue shift'.

You have already encountered the Doppler effect for sound waves (Chapter 15, Class XI textbook). For velocities small compared to the speed of light, we can use the same formulae which we use for sound waves. The fractional change in frequency $\Delta\nu/\nu$ is given by $-v_{\text{radial}}/c$, where v_{radial} is the component of the source velocity along the line joining the observer to the source. v_{radial} is considered positive when the source moves away from the observer. Thus, the Doppler shift can be expressed as:

$$\Delta\nu/\nu = -v_{\text{radial}}/c \quad (11.8)$$

The formula given above is valid only for speeds small compared to that of light. A more accurate formula for the Doppler effect which is valid even when the speeds are close to that of light, requires the use of Einstein's special theory of relativity. It is beyond the scope of this book. The Doppler effect for light is very important in astronomy. It is the basis for the measurements of the radial velocities of distant galaxies by Hubble and others which proved that the universe is expanding.

Example 11.1 What speed should a galaxy move with respect to us so that the sodium line at 589.0 nm is observed at 589.6 nm?

Answer Since $v\lambda = c$, $\frac{\Delta\nu}{\nu} = -\frac{\Delta\lambda}{\lambda}$ (for small changes in ν and λ),

$$\Delta\lambda = 589.6 - 589.0 = +0.6 \text{ nm}$$

Then, using Eq. (11.8), we get

$$\frac{\Delta\nu}{\nu} = -\frac{\Delta\lambda}{\lambda} = -v_{\text{radial}}/c$$

$$v_{\text{radial}} \approx +c \left(\frac{0.6}{589.0} \right) = +3.06 \times 10^5 \text{ m s}^{-1} \\ = 306 \text{ km/s.}$$

Therefore, the galaxy is moving away from us. ◀

11.3 COHERENT AND INCOHERENT ADDITION OF LIGHT WAVES

Our common experience is that when two light bulbs shine simultaneously on the same wall, the intensities add. If light consisted of particles, this would be easy to understand. An area of the wall receives the particles from both the sources, the number of particles striking a given area would be the sum of what each source would give by itself.

For waves from two different sources, the correct quantity to add is the displacement, not the intensity. You have already encountered this in the context of sound waves (Chapter 15, Class XI text book).

We recapitulate elementary concepts of superposition and interference using water waves. Imagine two sticks S_1 and S_2 being moved periodically up and down in an identical fashion in a trough of water (Fig. 11.6). They produce two water waves whose phase difference is stable, i.e., they are coherent. The oscillations in water level produced by the two sources will be in phase at any point P whose distances S_1P and S_2P from the two sources are equal. Graphs of displacement versus time as produced by each source separately are shown in Fig. 11.7(a). The lowermost graph shows the resultant displacement which is twice that of one source. This means that the intensity is four times than that produced by one source. This is called *constructive interference*.

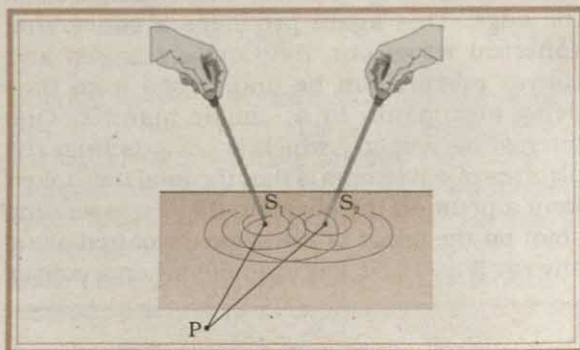


Fig. 11.6 Two sticks oscillating in phase in water, acting like two coherent sources.

Figure 11.7(c) shows a case when the path from one of the sources to a given point is half a wavelength longer than the other. This produces a phase difference of 180° or π radians between the two sources. Compare the first two graphs. The total displacement is zero in this case (when the amplitudes are equal). This situation is called *destructive interference*.

Figure 11.7(b) shows an intermediate case when the two oscillations have a phase difference of 90° . In this case, the resultant intensity of the combined oscillation is twice that due to one source and its phase is 45° , which is midway between the phases of two sources.

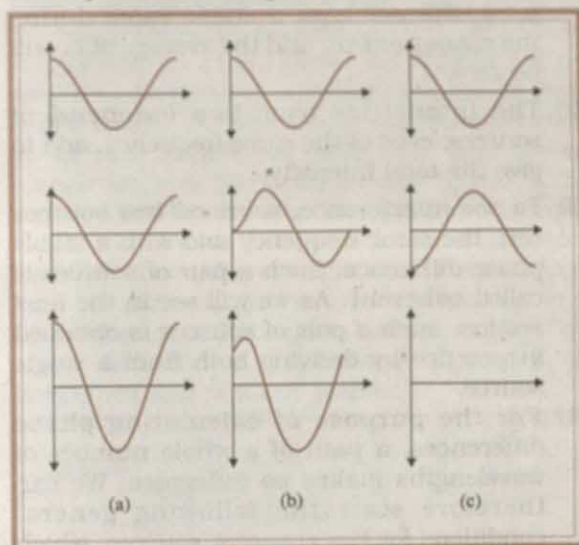


Fig. 11.7 Graphs of water level versus time. First row: as produced by the first source alone. Second row: as produced by the second source alone. Third row: combined effect of both sources (a) in phase, (b) phase difference of 90° and (c) phase difference of 180° .

Example 11.2 Combine two vibrations of equal amplitude and 90° phase difference.

Answer If we take the first vibration to be $\cos \omega t$, the second vibration is $\cos(\omega t - \pi/2) = \sin \omega t$. The superposition is

$$\begin{aligned} \sin \omega t + \cos \omega t &= \sin \omega t + \sin \left(\frac{\pi}{2} - \omega t \right) \\ &= 2 \sin \left[\frac{1}{2} \left(\omega t + \frac{\pi}{2} - \omega t \right) \right] \cos \left[\frac{1}{2} \left(\omega t + \omega t - \frac{\pi}{2} \right) \right] \\ &= 2 \sin \frac{\pi}{4} \cos \left(\omega t - \frac{\pi}{4} \right) = \sqrt{2} \cos \left(\omega t - \frac{\pi}{4} \right) \end{aligned}$$

The intensity is proportional to $(\sqrt{2})^2 = 2$, and the phase lags the first vibration by 45° .

For electromagnetic waves, the *electric fields* produced by different sources should be added.

The intensity I at a given point is proportional to the *square* of the electric field. For simplicity, we will assume that the angle between the two beams of light which fall at a given point from two sources is very small. We will also assume that the directions of the electric fields of the two waves are nearly the same. In this case, we need not worry about vector addition of electric fields and we can simply add the lengths of the two vectors (Fig. 11.8).

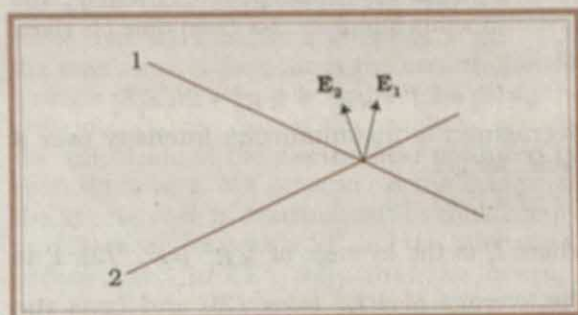


Fig. 11.8 Two rays 1 and 2 intersecting at a small angle, the electric vectors are nearly parallel.

We will see now that the *intensities* due to two sources *do not* add in general. Let the oscillating electric field produced by the first wave at the point P be given by $E_1 = a_1 \cos(\omega_1 t + \phi_1)$, and that produced by the second $E_2 = a_2 \cos(\omega_2 t + \phi_2)$. This means that the intensity produced by the first wave is equal to $k E_1^2 = k a_1^2 \cos^2(\omega_1 t + \phi_1)$, where k is a constant of proportionality which converts the square of the electric field into the intensity. Note that this expression oscillates very rapidly since the frequency of light is nearly 10^{15} Hz. What is observed in most experiments is the time average of this quantity. We note the following two properties which will help us to calculate such averages.

- The average of $\cos x$ is zero if the angle x varies uniformly over the full range 0 to 2π . Looking at the graph of $\cos x$, we can already guess that the positive and negative parts will cancel each other. Using $\cos(x \pm \pi) = -\cos x$, we prove that for every angle x , there is an angle $x + \pi$ [or $(x - \pi)$, if x is greater than π] for which the cosine is minus that for x . The average will clearly be zero.
- The average of $\cos^2 x$, when x varies uniformly over 0 to 2π radian, is $1/2$. To see this, we

write $\cos^2 x = \frac{1}{2} + \frac{1}{2} \cos 2x$ and use the earlier

result that the average of $\cos(2x)$ is zero.

The average intensity of the first wave will be given by $I_1 = k\alpha_1^2/2$. Similarly, the average of kE_2^2 over many cycles $I_2 = k\alpha_2^2/2$.

What is the average intensity when both the waves illuminate the same point and the electric fields are superposed? Clearly, the instantaneous intensity at a given time t is given by

$$k(E_1 + E_2)^2 = kE_1^2 + kE_2^2 + 2kE_1E_2$$

Averaging the instantaneous intensity over a cycle, we get

$$I = I_1 + I_2 + I_{12}$$

where I_1 is the average of kE_1^2 ($=k\alpha_1^2/2$), I_2 is the average of kE_2^2 ($=k\alpha_2^2/2$), and I_{12} is the average of $2kE_1E_2$. We have already seen the first two terms. They are the intensities which the individual sources would produce on their own. But this is not the end of the story. The total intensity I also contains the third term I_{12} . It is called the *interference term*. Let us see what this term is. I_{12} is the time average over a cycle of

$$2kE_1E_2 = 2k\alpha_1\alpha_2 \cos(\omega_1 t + \phi_1) \cos(\omega_2 t + \phi_2)$$

Using the trigonometric identity

$$2 \cos A \cos B = \cos(A+B) + \cos(A-B),$$

we get

$$2kE_1E_2 = k\alpha_1\alpha_2 [\cos\{(\omega_1 + \omega_2)t + (\phi_1 + \phi_2)\} + \cos\{(\omega_1 - \omega_2)t + (\phi_1 - \phi_2)\}] \quad (11.9)$$

Averaging $2kE_1E_2$ over many cycles, we find the first term (containing the sum of the two frequencies ω_1 and ω_2) becomes zero. The average of the second term is also zero if $\omega_1 \neq \omega_2$. The important conclusions are:

- (1) The interference term I_{12} averages to zero for two beams with *different* frequencies. This means that the time average intensities of two such beams can be added to give the total intensity.

Now, suppose the two frequencies are the same. We then have the average of $2kE_1E_2$ given by

$$I_{12} = k\alpha_1\alpha_2 \cos(\phi_1 - \phi_2) = \sqrt{I_1 I_2} \cos(\phi_1 - \phi_2) \quad (11.10)$$

The interference term is proportional to the cosine of the phase difference. However, when we speak of two different sources of light of the same frequency, say, two sodium lamps, the motions of the charges in them are independent. The phase difference $\phi_1 - \phi_2$ will not have a stable value during the measurement, and the average of I_{12} will be zero.

- (2) The intensities from two *independent* sources, even of the same frequency, add to give the total intensity.
- (3) To see interference, we need two sources with the same frequency *and* with a stable phase difference. Such a pair of sources is called **coherent**. As we will see in the next section, such a pair of sources is obtained in practice by deriving both from a single source.
- (4) For the purpose of calculating phase differences, a path of a whole number of wavelengths makes no difference. We can therefore state the following general conditions for two coherent sources which are themselves in phase.
 - (i) A path difference equal to an integer n times the wavelength gives constructive interference.
 - (ii) A path difference of half a wavelength plus an integer times the wavelength gives destructive interference.

The discussion just given helps us to understand why $I = I_1 + I_2$ is true in most experiments — the two sources have a phase difference which is not stable, but itself varies from 0 to 2π during the time interval of the measurement. The interference term then averages to zero. We call two such sources **incoherent**.

With sound waves, or at radio frequencies, one has much more control over the phases even of two independent sources and can make intensity measurements over short periods. For example, we get *beats* between two tuning forks even when the frequencies are different. There is no fundamental difference between electromagnetic waves with frequencies in the radio band and visible light. With modern techniques such as the use of lasers, it is possible to observe interference and beats between two independent sources even for visible light, but we do not consider these newer developments in this book.

11.4 INTERFERENCE

11.4.1 Young's Experiment

The experiment first carried out in the year 1802 by the English scientist Thomas Young is one of the most beautiful demonstrations of the wave nature of light. Two slits S_1 and S_2 are made in an opaque screen, parallel to each other and very close (Fig. 11.9). These are illuminated by another narrow slit S which is, in turn, lit by a bright source. As we will discuss in more detail in the next section on diffraction, light waves spread out from S and fall on both S_1 and S_2 . S_1 and S_2 then behave like two coherent sources — the two sticks in our water wave example. Thus, the two coherent light waves are derived from the same original source. In this way, any phase change in S occurs in both S_1 and S_2 . The phase difference $\phi_1 - \phi_2$ between S_1 and S_2 is unaffected and remains stable.

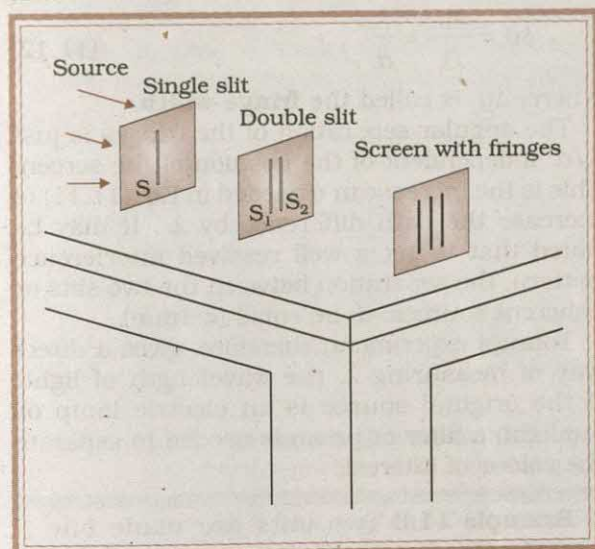


Fig. 11.9 A schematic view of Young's double slit interference experiment. The widths of the slits, their separation, and the fringes are not to scale. They have been shown bigger for clarity.

Light now spreads out from both S_1 and S_2 and falls on a screen. It is essential that the waves from the two sources overlap on the same part of the screen. If one slit is covered up, the other produces a wide smoothly illuminated patch on the screen (which we will study in detail in the section on diffraction). But when both the slits are open, the patch is seen to be crossed by dark and bright bands called *interference fringes*. A part of the screen which receives light

from either S_1 or S_2 alone can become dark when both slits are open! This is a clear case of destructive interference and hence a proof of the wave nature of light.

Figure 11.10 shows graphs of how the intensity on the screen varies as one moves along a line perpendicular to the slits. The thick line shows the case when both slits are open and the fringes are formed. For comparison, the intensity received at the screen from a single slit is also shown as a dotted line. These graphs show that the minimum intensity is zero and the maximum is four times the contribution of a single slit. This is easily understood using the principle of superposition. For simplicity, we take the amplitude of the electric field produced by each slit to be E . We get zero electric field when the interference is destructive. For constructive interference, the field is $2E$ and the intensity is proportional to $4E^2$. Note that the *average* intensity of the maximum and minimum is $2E^2$, just the sum of what the two slits produce. The phenomenon of interference, therefore, redistributes the energy keeping the total intensity constant.

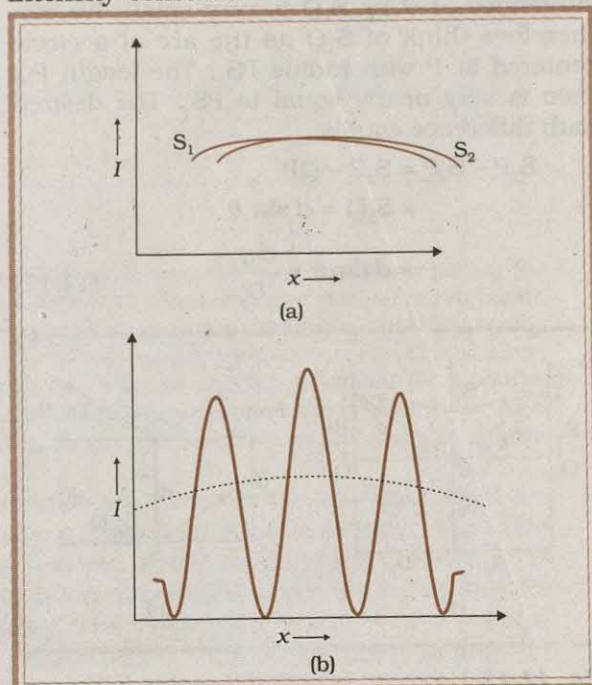


Fig. 11.10 Intensity I of light plotted against position x on the screen in Young's experiment: (a) Slit S_1 or slit S_2 individually opened. (b) Both slits open showing interference fringes. The dashed curve is the single slit intensity for comparison.

11.4.2 Path Differences

To understand the fringe pattern in more detail, we need to calculate the path lengths between the slits S_1 and S_2 , and points on the screen. Fig. 11.11(a) shows three parallel planes. The first contains the slit S , the second the two slits S_1 and S_2 , and the third a typical point P where the light falls on the screen. The point O is midway between S_1 and S_2 . The perpendicular to all the planes through O meets the source plane at O_2 and the screen at O_1 . These two points are used as origins to measure the positions of S and P .

$$O_1P = y_1; \quad O_2S = y_2$$

To start with, assume that S is equidistant from S_1 and S_2 [that is, S coincides with O_2 so that the angle ϕ in Fig. 11.11(a) is zero]. We need to calculate the path difference, $S_2P - S_1P$. As shown in Fig. 11.11(b), the two lines S_1P and S_2P are nearly parallel since the distance $S_1S_2 = d$ is much less than $OO_1 = D_1$. The angle that these two lines make with the normal to the screen is denoted by θ . A perpendicular S_1Q is dropped from S_1 to the line S_2P . The angle subtended at P by S_1Q is very small. We can therefore think of S_1Q as the arc of a circle centered at P with radius PS_1 . The length PQ then is very nearly equal to PS_1 . The desired path difference equals

$$\begin{aligned} S_2P - S_1P &= S_2P - QP \\ &= S_2Q = d \sin \theta \\ &\approx d \tan \theta = \frac{dy_1}{D_1} \end{aligned} \quad (11.11)$$

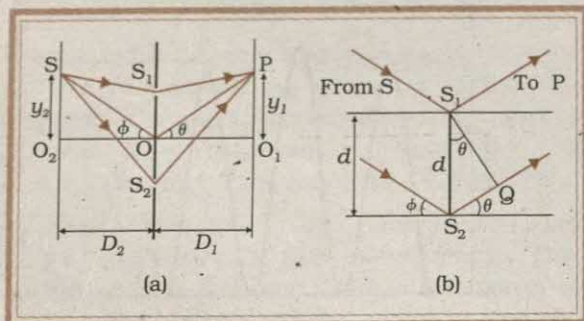


Fig. 11.11 Geometry of path differences in Young's experiment. (a) Coordinate system defining the positions of the slit S , double slit S_1S_2 and point of observation on the screen, P . (b) Magnified view of the paths reaching the two slits from the source S and travelling to the screen at P .

In deriving Eq. (11.11) we have used the result for small angles that $\sin \theta \approx \theta \approx \tan \theta$ (for small θ). The condition for constructive interference reads

$$\begin{aligned} S_2P - S_1P &= \frac{dy_1}{D_1} = n\lambda \\ \text{i.e., } y_1 &= n \frac{D_1\lambda}{d} \end{aligned}$$

Putting $n = 0$, we have a central bright fringe corresponding to zero path difference at $y_1 = 0$. The separation between two successive maxima is found by subtracting the values of y_1 corresponding to successive values of n . We denote this by Δy_1

$$\Delta y_1 = \frac{D_1\lambda}{d} (n+1 - n) = \frac{D_1\lambda}{d}$$

This separation Δy_1 between two adjacent bright (or dark) fringes subtends an angle $\Delta\theta$ at the centre O of the double slit.

$$\Delta\theta = \frac{\Delta y_1}{D_1} = \frac{\lambda}{d} \quad (11.12)$$

where, Δy_1 is called the **fringe width**.

The angular separation of the fringes is just λ/d independent of the position of the screen. This is the increase in θ needed in Eq. (11.11) to increase the path difference by λ . It may be noted that to get a well resolved interference pattern, the separation between the two slits or coherent sources, d , be small ($< 1\text{ mm}$).

Young's experiment, therefore, gives a direct way of measuring λ , the wavelength of light. If the original source is an electric lamp or sunlight, a filter or prism is needed to separate the colour of interest.

Example 11.3 Two slits are made one millimetre apart and the screen is placed one metre away. What is the fringe separation when blue-green light of wavelength 500 nm is used?

$$\begin{aligned} \text{Answer} \quad \text{Fringe spacing} &= \frac{D\lambda}{d} = \frac{1 \times 5 \times 10^{-7}}{1 \times 10^{-3}} \text{ m} \\ &= 5 \times 10^{-4} \text{ m} = 0.5 \text{ mm} \end{aligned}$$

For simplicity, we have so far placed the first slit S at $y_2 = 0$. This produced zero path difference at $y_1 = 0$. When we have a source at y_2 , the lines SS_1 and SS_2 make an angle $\phi = y_2/D_2$ with the x -axis [Fig. 11.11(b)]. At the central fringe, the total path difference is zero,

$$SS_2 + S_2P - SS_1 - S_1P = (S_2P - S_1P) + (S_2S - S_1S) = 0;$$

$$\frac{dy_2}{D_2} + \frac{dy_1}{D_1} = 0$$

$$\frac{y_2}{D_2} = -\frac{y_1}{D_1} \text{ or } \theta = -\phi \quad (11.13)$$

The meaning of Eq. (11.13) is that the angular movement θ of the central fringe is equal and opposite to the angular movement ϕ of the source. Both the angles are measured at the centre of the double slit. More simply, the first slit, the centre of the double slit and the central fringe lie in a straight line as the source is moved.

VISIBILITY OF THE INTERFERENCE FRINGES

For a basic course such as this one, it is sufficient to consider the simplest case of a monochromatic point source illuminating both the slits in Young's experiment. However, this box explains what happens as we allow for the real situation in which the source has a finite size and radiates wavelengths in a finite range.

Suppose one uses a source (like the filament of a bulb) in place of the narrow slit S . Let this have a size Δy_2 . Each part of the source produces a fringe pattern on the screen. Since these different parts are incoherent (independent sources), we can add the intensities of the different fringe patterns on the screen. Their central fringes will be spread out over an angle $\Delta\theta = \Delta y_2/D_2$. The effect of adding just two equal fringe patterns is shown in Fig. 11.12(a). If the maximum of one falls on the minimum of the other, then there is no intensity variation in the total. This means the fringes are no longer visible.

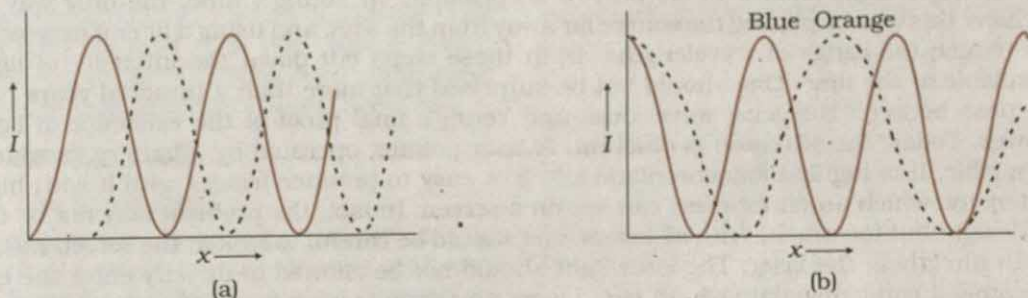


Fig. 11.12 Two factors affecting the visibility of interference fringes. (a) Each incoherent part of the source produces its own fringe pattern. Because their maxima and minima are different, the fringe contrast is poor. (b) Different wavelengths present in the source add incoherently, each producing its own fringe pattern. The longer wavelengths like orange (dashed line) produce the dashed fringe pattern which has a larger spacing than that for a shorter wavelength like blue (solid line). The different patterns agree at the central fringe (zero path difference). At larger path differences, the fringes become coloured and then fainter.

If the angle $\Delta\theta$ is much less than λ/d (the angular separation of two maxima), then the fringes from the different parts of the source have maxima and minima at nearly the same place on the screen. We then see the full fringe pattern in the total intensity. But if $\Delta\theta$ is much greater than λ/d , then the maxima of the different patterns which we are adding are spread out by much more than the fringe spacing. The fringe pattern will be washed out. In Example 11.3, the spacing between bright fringes was 0.5 mm subtending an angle of $(1/2000)$ rad at the slits. Any source subtending a much larger angle than 5×10^{-4} rad, i.e., $100''$, at the slits will not give clear fringes. For comparison, the Sun subtends approximately $(\frac{1}{2})^\circ = 1800''$.

We now allow for the presence of a range of wavelengths in the incident light covering a range $\Delta\lambda$ around λ . When we speak of a monochromatic source, it just means $\Delta\lambda$ is small compared to λ , not that $\Delta\lambda$ is zero! The central fringe $n = 0$ occurs at $\theta = 0$ for all wavelengths. But take the example of the $n = 2$ fringe for blue light of $\lambda = 450$ nm. The path difference is

900 nm. For orange light of $\lambda = 600$ nm, this path difference is one and a half wavelengths and there is destructive interference. Figure 11.12(b) shows two fringe patterns for different wavelengths. We add the intensities since different wavelengths are incoherent. The fringes near the centre agree for all wavelengths. As the path difference increases, the fringes get washed out because the maximum for one wavelength happens to be a minimum for some other wavelength. By making the range of wavelengths $\Delta\lambda$ narrower, one can get a larger number of fringes. For example, let the range be 450 nm to 500 nm. It is only for $n = 5$ for

$\lambda = 450$ nm that we have $n = 4\frac{1}{2}$ for $\lambda = 500$ nm. This means that about 5 fringes are seen distinctly on either side of the central maximum.

To summarise, two slits separated by d and illuminated coherently by a source of angular size $\Delta\theta$, give fringes only if

$$\Delta\theta \leq \frac{\lambda}{d}$$

Further, the n^{th} fringe will be visible only if the range of wavelength, $\Delta\lambda$, satisfies $n < \frac{\lambda}{\Delta\lambda}$

i.e.,
$$\Delta\lambda < \frac{\lambda}{n}$$

The conditions for clearly visible fringes in Young's experiment are that the source should subtend as small an angle at the two slits, and that the range of wavelengths present in the incident light should be as narrow as possible. In Young's time, the only way to achieve this was by placing the source far away from the slits, and using a filter if necessary to reduce the range of wavelengths. Both these steps cut down the intensity of light available at the time. One should not be surprised that more than a hundred years had to pass between Huygens' wave ideas and Young's final proof of the existence of light waves. Today, the situation is different. A laser pointer, operated by a battery, is widely available. It is highly monochromatic and it is easy to produce fringes with it with high intensity, which an entire class can see on a screen. In fact, the problem may not be too little light but too much. With all lasers, one should be careful to look at the screen rather than directly at the laser. The laser light should not be allowed to directly enter the eye because it can cause damage. In fact, lasers are actually used for surgery on the eye.

11.4.3 The Colours of Thin Films

During the rainy season, one sometimes sees oil from some motor vehicles on the road, spreading out to form a thin layer on water. Such a layer often shows brilliant colours even when illuminated with white light. Another thin film showing brilliant colours is a soap bubble. In this case, the film slowly becomes thinner near the top of the bubble as the water slowly moves down towards the bottom of the bubble. It is then found that this very thin layer near the top appears dark in reflected light. These colours shown by thin films can be understood as interference between two beams. One is reflected from the top surface of the film and the other from the bottom surface (Fig. 11.13). Since they both originate from the same source, they are coherent.

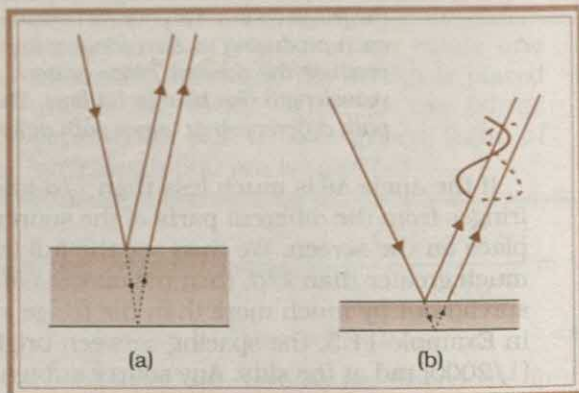


Fig. 11.13 Thin film interference. (a) General case. (b) Thickness much less than wavelength, showing destructive interference.

As an example, consider a path difference of 1000 nm. This leads to destructive interference

(two and a half wavelengths) in the violet side of the spectrum $\lambda = 400\text{nm}$, but constructive interference (two wavelengths) in the green region ($\lambda = 500\text{ nm}$). The reflected light will thus be coloured. The connection between film thickness and colour was studied experimentally by Newton. However, he could not explain it satisfactorily as he did not fully believe in wave ideas.

One might think that a very thin film would give zero path difference and hence constructive interference at all wavelengths. But what is observed, for example, in a thin soap film, is actually destructive interference. Actually there is an additional phase difference of π between a ray in air (rarer medium) reflected from water (denser medium), and a ray travelling in water reflected from air. This result explains why there is destructive interference for a very thin film. This is also what we should expect from a different argument. For a very thin film, the number of molecules available to scatter the light back is very small, so such a film must look dark in reflected light. This extra phase change of π is used in the problems section, where many examples of thin film interference are given.

11.5 DIFFRACTION

11.5.1 The Single Slit

In the discussion of Young's experiment, we stated that a single narrow slit acts as a new source from which light spreads out. Even before Young, early experimenters — including Newton — had noticed that light spreads out from narrow holes and slits. It seems to turn around corners and enter regions where we would expect a shadow. These effects, known as *diffraction*, can only be properly understood using wave ideas. After all, you are hardly surprised to hear sound waves from someone talking around a corner!

When the double slit in Young's experiment is replaced by a single narrow slit (illuminated by a monochromatic source), a broad pattern with a central bright region is seen. On both sides, there are alternate dark and bright regions, the intensity becoming weaker away from the centre (Fig. 11.15). To understand this, go to Fig. 11.14, which shows a parallel beam of light falling normally on a single slit LN of width a . The diffracted light goes on to meet a screen. The midpoint of the slit is M.

A straight line through M perpendicular to the slit plane meets the screen at C. We want the intensity at any point P on the screen. As before, straight lines joining P to the different points L, M, N, etc., can be treated as parallel, making an angle θ with the normal MC.

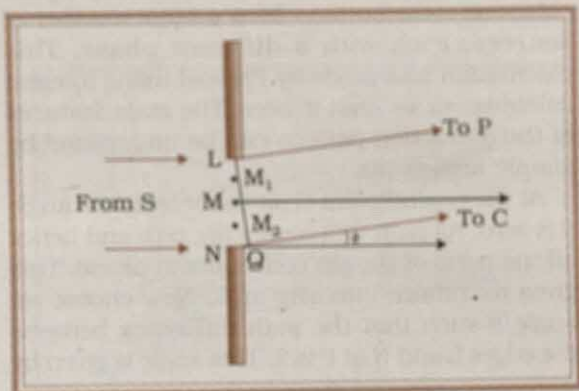


Fig. 11.14 The geometry of path differences for diffraction by a single slit.

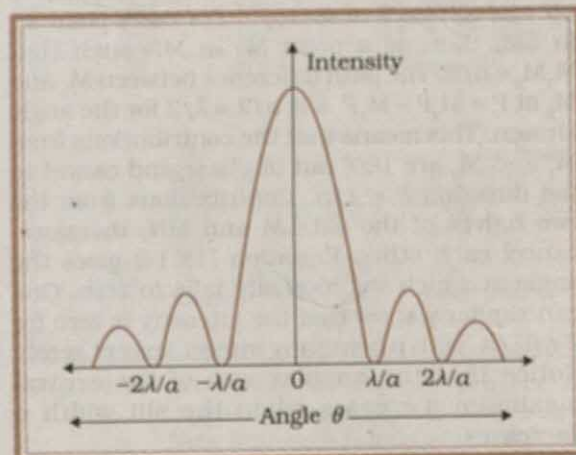


Fig. 11.15 The variation of intensity with angle in single slit diffraction. The first secondary maximum is only 4% of the central maximum, so it is not to scale in the figure.

The basic idea is to divide the slit into much smaller parts, and add their contributions at P with the proper phase differences. We are treating different parts of the wavefront at the slit as secondary sources. Because the incoming wavefront is parallel to the plane of the slit, these sources are in phase.

The path difference NP – LP between the two edges of the slit can be calculated exactly as for Young's experiment. From Fig. 11.14,

$$\begin{aligned} NP - LP &= NQ \\ &= a \sin \theta \\ &\approx a\theta \end{aligned}$$

Similarly, if two points M_1 and M_2 in the slit plane are separated by y , the path difference $M_2P - M_1P \approx y\theta$. We now have to sum up equal, coherent contributions from a large number of sources, each with a different phase. This calculation was made by Fresnel using integral calculus, so we omit it here. The main features of the diffraction pattern can be understood by simple arguments.

At the central point C on the screen, the angle θ is zero. All path differences are zero and hence all the parts of the slit contribute in phase. This gives maximum intensity at C. Now choose an angle θ such that the path difference between the edges L and N at P is λ . This angle is given by

$$a\theta = \lambda, \quad \theta = \frac{\lambda}{a}. \quad (11.14)$$

Further, divide the slit into two equal halves LM and MN each of size $a/2$. For every point M_1 in LM, there is a point M_2 in MN such that $M_1M_2 = a/2$. The path difference between M_1 and M_2 at P = $M_2P - M_1P = \theta \cdot a/2 = \lambda/2$ for the angle chosen. This means that the contributions from M_1 and M_2 are 180° out of phase and cancel in the direction $\theta = \lambda/a$. Contributions from the two halves of the slit LM and MN, therefore, cancel each other. Equation (11.14) gives the angle at which the intensity falls to zero. One can similarly show that the intensity is zero for $\theta = n\lambda/a$, with n being any integer (except zero!). Notice that the angular size of the central maximum increases when the slit width a decreases.

Let us now consider an angle $\theta = 3\lambda/2a$ which is midway between two of the dark fringes $\theta_1 = \lambda/a$ and $\theta_2 = 2\lambda/a$. If we take the first two thirds of the slit, the path difference between the two ends would be

$$\frac{2}{3}a \times \theta = \frac{2a}{3} \times \frac{3\lambda}{2a} = \lambda \quad (11.15)$$

The first two-thirds of the slit can therefore be divided into two halves which have a $\lambda/2$ path difference. The contributions of these two halves cancel in the same manner as described earlier. Only the remaining one-third of the slit contributes to the intensity at a point between the two minima. Clearly, this will be much

weaker than the central maximum (where the entire slit contributes in phase). One can similarly show that there are maxima at $(n + 1/2)\lambda/a$ with $n = 2, 3$, etc. These become weaker with increasing n , since only one-fifth, one-seventh, etc., of the slit contributes in these cases.

The graph in Fig. 11.15 shows how the intensity on the screen varies with the angle θ . It shows the central maximum at $\theta = 0$, zero intensity at $\theta = \pm n\lambda/a$ ($n \neq 0$) and secondary maxima at $\theta = \pm (n + 1/2)\lambda/a$ (for $n \neq 0$).

We now compare and contrast the pattern which is seen with two coherently illuminated narrow slits in Young's experiment (usually called an interference pattern) with that seen for a coherently illuminated single slit (usually called the single slit diffraction pattern).

- (1) The interference pattern has a number of equally spaced bright and dark bands of equal intensity. The diffraction pattern has a central bright maximum which is twice as wide as the other maxima. The intensity falls as we go to successive maxima away from the centre, on either side.
- (2) We calculate the interference pattern by superposing two waves originating from the two narrow slits. The diffraction pattern is a superposition of a continuous family of waves originating from each point on a single slit.
- (3) For a single slit of width a , the first null of the interference pattern occurs at an angle of λ/a . At the same angle of λ/a , we get a maximum (not a null) for two narrow slits separated by a distance a .

Having pointed out these differences, we should emphasise that both the patterns are governed by the same ideas of superposition and Huygens' secondary wavelets. In a realistic Young experiment, since the individual slits will produce their own diffraction patterns, one would observe both interference as well as diffraction patterns.

We now go back to the intensity produced by each slit in Young's experiment when the other is blocked (Fig. 11.15). We have a central maximum which is broad if the slit is made narrow. The graphs given in Fig. 11.15 show only a small portion of the central maximum of the single slit diffraction pattern.

Example 11.4 In Example 11.3, what should the width of each slit be to obtain 10 maxima of the double slit pattern within the central maximum of the single slit pattern?

Answer We want

$$a\theta = \lambda, \theta = \frac{\lambda}{a} \quad (11.14)$$

$$10 \frac{\lambda}{d} = 2 \frac{\lambda}{a}; \quad a = \frac{d}{5} = 0.2 \text{ mm}$$

Notice that the wavelength of light and distance of the screen do not enter in the calculation of a .

In our discussion of Young's experiment and the single slit diffraction, we have assumed that the screen on which the fringes are formed is at a large distance. The two or more paths from the slits to the screen were treated as parallel. This situation also occurs when we place a converging lens after the slits and place the screen at the focus. Parallel paths from the slit are combined at a single point on the screen [Fig. 11.16(a)]. Note that the lens does not introduce any extra path differences in a parallel beam, as we remarked in Section 11.2. This arrangement is often used since it gives more intensity than placing the screen far away. If f is the focal length of the lens, then we can easily work out the size of the central bright maximum. In terms of angles, the separation of the central maximum from the first null of the diffraction pattern is λ/a . Hence, the size on the screen will be $f\lambda/a$.

11.5.2 Resolving Power of Telescopes and Microscopes

We will now see that diffraction plays a fundamental role in determining the ability of telescopes and microscopes to perceive fine details in the object.

When a light wave enters an optical instrument with a circular opening (objective lens of diameter D), a diffraction pattern is produced by the interference of the wavelets originating at different points in the aperture [see the first minimum for a Fig. 11.16(a)]. According to ray optics, the image of a point source formed by an ideal lens is a point. However, because of the diffraction effects, the image of a single point in the object formed by a

circular lens will actually be a bright central circular region surrounded by concentric dark and light rings [Fig. 11.16(b)]. A detailed analysis shows that the first minimum for a circular opening of diameter D occurs when

$$\sin \theta = \theta = 1.22 \frac{\lambda}{D} \quad (11.16)$$

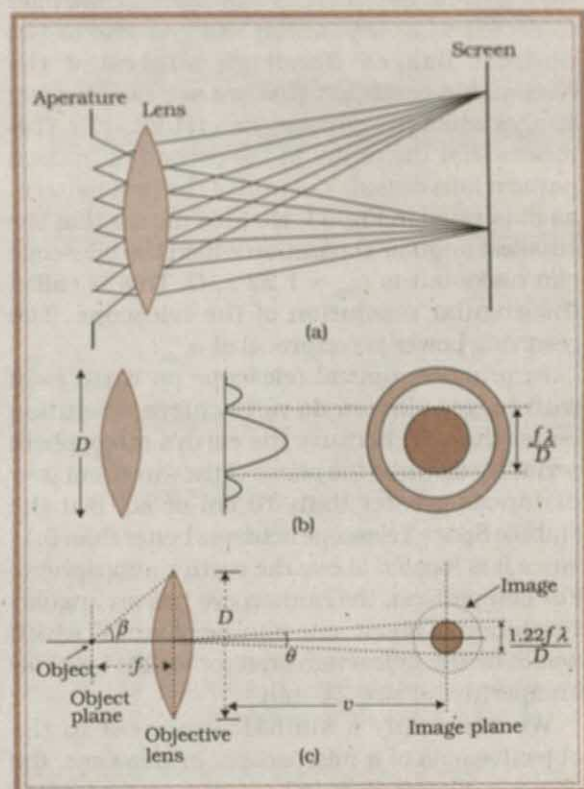


Fig. 11.16 (a) Using a lens to project the diffraction pattern of a slit on a screen. Note that each family of parallel rays reaches one point on the screen. (b) Intensity distribution at the focus of a lens. Note the central spot. (c) Real image formed by the objective lens of a microscope.

This relation is similar to the formula for the first minimum of a single slit, $\sin \theta = \lambda/a$, but the circular geometry results in the factor 1.22.

As discussed in the earlier sub-section, if a lens of focal length f is used to focus the diffraction, the size of the central maximum on the screen will be

$$f\theta = f \left(\frac{1.22 \lambda}{D} \right) \quad (11.17)$$

Consider a telescope looking at two stars separated by an angle α in the sky. Each one produces a diffraction pattern as image, and the centres of these two patterns are separated by $f\alpha$. If α is very small, this separation will be less than the size of each pattern. The consequence would be the overlapping of the two patterns and we will think that there is only one star. We have to decide what separation will give rise to two distinct images. Rayleigh suggested the reasonable condition that we see two distinct images when $f\alpha$ is greater than $f(1.22 \lambda/D)$. This means that the centre of the second diffraction pattern falls outside the null of the first pattern, as illustrated in Fig. 11.15. This means that the smallest angular separation which the telescope can make out is $\alpha_{\min} = 1.22 \lambda/D$. This is called the angular resolution of the telescope. The resolving power is reciprocal of α_{\min} .

In practice, optical telescope on earth even with bigger mirrors do not achieve resolution better than $1''$, because the earth's atmosphere seriously disturbs the phase of the wavefront over distances greater than 10 cm or so. But the Hubble Space Telescope achieves better than $0.1''$ since it is located above the earth's atmosphere. For comparison, the human eye has an angular resolution of about one minute of arc ($1'$) which would be the diffraction limit for visible light, for an aperture of size 1.6 mm.

We can apply a similar argument to the objective lens of a microscope. In this case, the object is placed slightly beyond f , so that a real image is formed at a distance v [Fig. 11.16(c)]. The magnification — ratio of image size to object size — is given by $m = v/f$. It can be seen from the Fig. 11.16(c) that

$$D/f = 2 \tan \beta \quad (11.18)$$

where 2β is the angle subtended by the diameter of the objective lens at the focus of the microscope.

When the separation between two points in a microscopic specimen is comparable to the wavelength λ of the light, the diffraction effects become important. The image of a point object will again be a diffraction pattern whose size in the image plane will be

$$v\theta = v \left(\frac{1.22\lambda}{D} \right) \quad (11.19)$$

Two objects whose images are closer than this distance will not be resolved, they will be seen as one. The corresponding minimum separation, d_{\min} , in the object plane is given by

$$\begin{aligned} d_{\min} &= \left[v \left(\frac{1.22\lambda}{D} \right) \right] / m \\ &= \frac{1.22\lambda}{D} \cdot \frac{v}{m} \\ &= \frac{1.22 f \lambda}{D} \end{aligned} \quad (11.20)$$

Now, combining Eqs. (11.18) and (11.20) we get

$$\begin{aligned} d_{\min} &= \frac{1.22\lambda}{2 \tan \beta} \\ &\approx \frac{1.22\lambda}{2 \sin \beta} \end{aligned} \quad (11.21)$$

If the medium between the object and the objective lens is not air but a medium of refractive index n , Eq. (11.21) gets modified to

$$d_{\min} = \frac{1.22\lambda}{2n \sin \beta} \quad (11.22)$$

The product $n \sin \beta$ is called the *numerical aperture* and is sometimes marked on the objective.

The resolving power of the microscope is given by the reciprocal of the minimum separation of two points seen as distinct. It can be seen from Eq. (11.22) that the resolving power can be increased by choosing a medium of higher refractive index. Usually an oil having a refractive index close to that of the objective glass is used. Such an arrangement is called an '*oil immersion objective*'. Notice that it is not possible to make $\sin \beta$ larger than unity. Thus, we see that the resolving power of a microscope is basically determined by the wavelength of the light used. You will see later (Chapter 12) that higher resolving power is achieved by using electrons which behave as waves.

Example 11.5 What is the angular resolution of a 10 cm diameter telescope at a wavelength of $0.5 \mu\text{m}$?

$$\begin{aligned} \text{Answer } \Delta\theta &= \frac{1.22\lambda}{D} = \frac{0.6 \times 10^{-6}}{0.1} \\ &= 6 \times 10^{-6} \text{ rad} = 1.2'' \end{aligned}$$

11.5.3 Seeing the Single Slit Diffraction Pattern

It is surprisingly easy to see the single slit diffraction pattern for oneself. The equipment needed can be found in most homes — two razor blades and one clear glass electric bulb preferably with a straight filament. One has to hold the two blades so that the edges are parallel and have a narrow slit in between. This is easily done with the thumb and forefingers (Fig. 11.17). Keep the slit, parallel to the filament, right in front of the eye. Use spectacles if you normally do. With slight adjustment of the width of the slit and the parallelism of the edges, the pattern should be seen with its bright and dark bands. Since the position of all the bands (except the central one) depends on wavelength, they will show some colours. Using a filter for red or blue will make the fringes clearer. With both filters available, the wider fringes for red compared to blue can be seen.

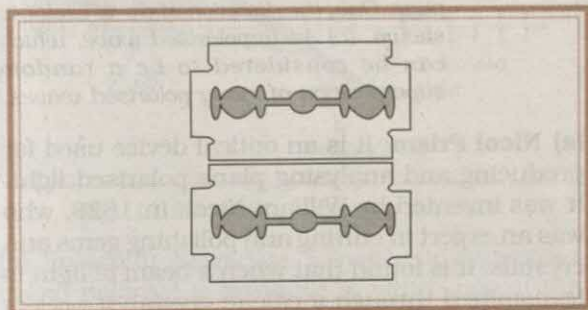


Fig. 11.17 Holding two blades to form a single slit. A bulb filament viewed through this shows clear diffraction bands.

In this experiment, the filament plays the role of the first slit S in Fig. 11.9. The lens of the eye focuses the pattern on the screen (the retina of the eye).

With some effort, one can cut a double slit in an aluminium foil with a blade. The bulb filament can be viewed as before to repeat Young's experiment. In daytime, there is another suitable bright source subtending a small angle at the eye. This is the reflection of the Sun in any shiny convex surface (e.g., a cycle bell). Do not try direct sunlight — it can damage the eye and will not give fringes anyway as the Sun subtends an angle of $(1/2)^\circ$.

11.5.4 The Validity of Ray Optics

An aperture (i.e., slit or hole) of size a illuminated by a parallel beam sends diffracted light into an

angle of approximately $\approx \lambda/a$. This is the angular size of the bright central maximum. In travelling a distance z , the diffracted beam therefore acquires a width $(z \lambda/a)$ just due to diffraction. It is interesting to ask at what value of z the spreading due to diffraction becomes greater than the size a of the aperture. Figure 11.18 shows how the nature of the beam changes from parallel to divergent when this happens. We want

$$\frac{2z\lambda}{a} > a, \text{ i.e., } z > \frac{a^2}{2\lambda} \quad (11.23)$$

We define a quantity z_F called the *Fresnel distance* by the following equation

$$z_F = a^2 / \lambda$$

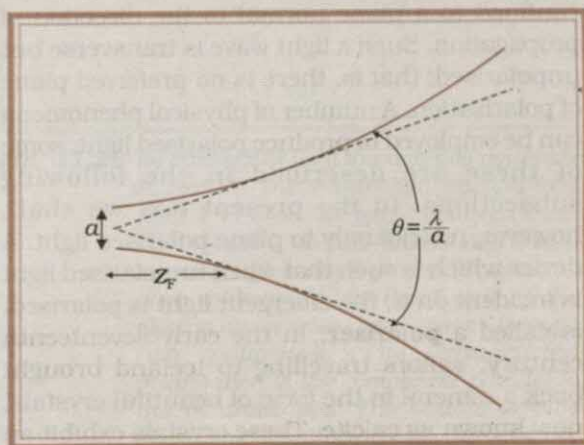


Fig. 11.18 A parallel beam of size ' a ' starts to diverge appreciably after travelling a distance z_F .

Eq. (11.23) above shows that for distances much smaller than z_F , the spreading due to diffraction is smaller compared to the size of the beam. It becomes comparable when the distance is $z_F/2$. For distances much greater than z_F , the spreading due to diffraction dominates over that due to ray optics (i.e., the size a of the aperture).

Example 11.6 For what distance is ray optics a good approximation when the aperture is 3 mm wide and the wavelength 500 nm?

Answer $z_F = \frac{a^2}{\lambda} = \frac{(3 \times 10^{-3})^2}{5 \times 10^{-7}} = 18 \text{ m}$

This example shows that even with a small aperture, diffraction spreading can be neglected for rays many metres in length. Thus, ray optics is valid in many common situations. ◀

11.6 POLARISATION

In Chapter 9, we have learnt that light consists of transverse waves in which the electric vector is confined to a plane perpendicular to the direction of propagation of the wave. The wave illustrated in Fig. 9.6 in which the electric vector is confined to the x - y plane, is said to be *linearly polarised* (also called plane polarised). Generally, in light obtained from a source such as the Sun or an incandescent lamp, the orientation of electric vector is not confined to one plane only. However, when viewed from along the direction of propagation, the electric vector appears to be randomly oriented [Fig. (11.19)], although confined to a plane normal to the direction of propagation. Such a light wave is transverse but unpolarised; that is, there is no preferred plane of polarisation. A number of physical phenomena can be employed to produce polarised light, some of these are described in the following subsections. In the present text we shall, however, restrict only to plane polarised light. A device which is such that when unpolarised light is incident on it, the emergent light is polarised, is called a **polariser**. In the early seventeenth century, sailors travelling to Iceland brought back a mineral in the form of beautiful crystals, now known as calcite. These crystals exhibit an interesting phenomenon called double refraction. It was found that a ray of light entering a calcite crystal splits into two rays which travel with different speeds, i.e., have different refractive indices. The reason for this behaviour was later understood to be that the electric vectors of these two rays are in two perpendicular directions. A suitable combination of such crystals can be employed to design a polariser. Nowadays a synthetic substance called **polaroid** is available which produces the same effect. We shall discuss more about these two types of polarisers in the subsequent section. In principle, all polarisers can be used as analysers for the polarised light. This is illustrated in reference to polaroids.

It may be noted that phenomena of interference and diffraction illustrate the wave nature of light. However, the transverse nature of light waves is demonstrated only by the phenomenon of polarisation.

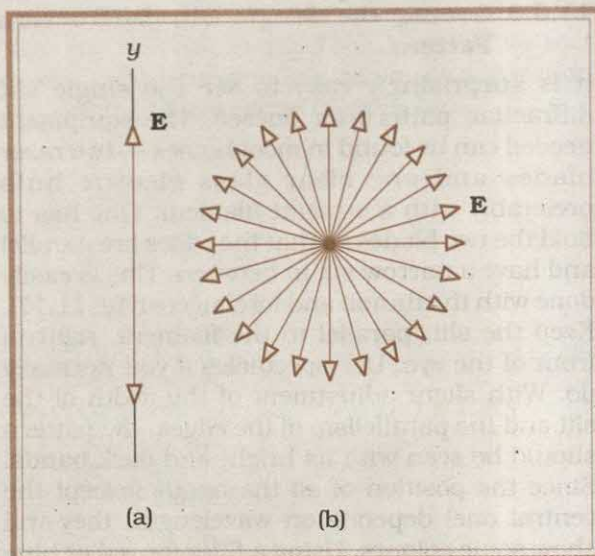


Fig. 11.19 (a) A linearly polarised wave, viewed from along the direction of propagation. The wave is moving out of the plane of the page. Only the direction of the E vector is shown. (b) An unpolarised wave, which can be considered to be a random superposition of many polarised waves.

(a) Nicol Prism: It is an optical device used for producing and analysing plane polarised light. It was invented by William Nicol, in 1828, who was an expert in cutting and polishing gems and crystals. It is found that when a beam of light is transmitted through a calcite crystal, it breaks up into two rays: (1) the ordinary ray which has its electric vector perpendicular to the principal section of the crystal and (2) the extraordinary ray which has its electric vector parallel to the principal section.

The nicol prism (Fig. 11.20) is made in such a way that it eliminates one of the two rays by total internal reflection. It is generally found that the ordinary ray is eliminated and only the extraordinary ray is transmitted through the prism. The nicol prism consists of two calcite crystals cut at $\sim 68^\circ$ with its principal axis joined by a glue called Canada balsam. Canada balsam has a refractive index of 1.55 while the refractive indices of calcite for the ordinary and extraordinary rays are, respectively, 1.658 and 1.486.

In Fig. 11.20, the principal section ACGE of the crystal is shown. The diagonal AG represents the Canada balsam layer. The refractive index

for the ordinary ray is more than that for the extraordinary ray. The refractive index of Canada balsam lies between the refractive indices of calcite for the ordinary and the extraordinary rays.

Thus, Canada balsam acts as a rarer medium for an ordinary ray and it acts as a denser medium for the extraordinary ray. Therefore, when the ordinary ray passes from a portion of the crystal into the layer of Canada balsam, it passes from a denser to a rarer medium. When the angle of incidence is greater than the critical angle i_c ($= 69^\circ$), the ray is totally internally reflected and is not transmitted. The extraordinary ray is not affected and is, therefore, transmitted through the nicol prism. Therefore, a ray of unpolarised light on passing through the nicol prism (Fig. 11.20) becomes plane-polarised.

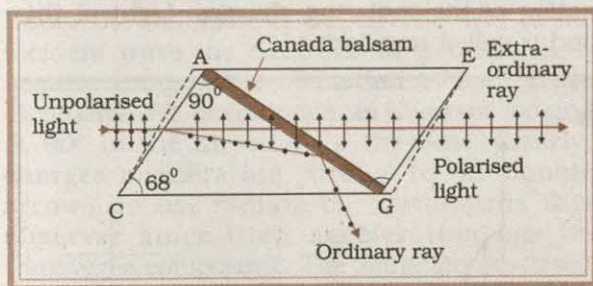


Fig. 11.20 Nicol prism as a polariser.

(b) Polaroid: Nowadays a synthetic substance called polaroid is available in which one of the rays, with a particular direction of the electric vector along some dye molecules, is absorbed much more than the other. For practical purposes, a polaroid can be said to transmit waves having only one specific direction of the electric vector.

When light from a bulb passes through a single piece of polaroid P_1 , it is observed that its intensity is cut down to half [Fig. 11.21(a)]. Rotating P_1 seems to have no effect on the transmitted beam since the transmitted intensity remains constant. Now, let an identical piece polaroid P_2 be placed before P_1 . As expected, the light from the bulb is reduced in intensity on passing through P_2 alone. But now rotating P_1 has a dramatic effect on the light coming from P_2 . In one position, the intensity transmitted by P_2 followed by P_1 is nearly zero. When turned by 90° from this position, P_1 transmits nearly the full intensity emerging from P_2 [Fig. 11.21(a)].

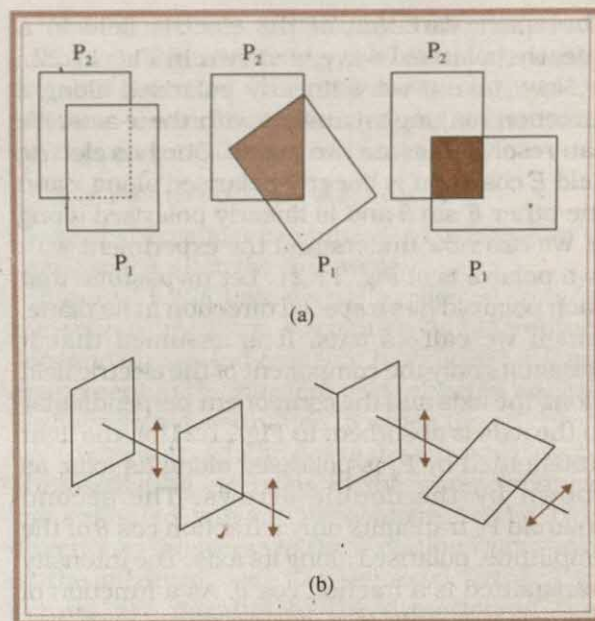


Fig. 11.21 (a) Passage of light through two polaroids P_2 and P_1 . The transmitted fraction falls from 1 to 0 as the angle between them varies from 0° to 90° . Notice that the light seen through a single polaroid P_1 does not vary with angle. (b) Behaviour of the electric vector when light passes through two polaroids. The transmitted polarisation is the component parallel to the polaroid axis. The double arrows show the oscillations of the electric vector.

Clearly, the light transmitted by P_2 has some special property, which the light from the bulb does not have.

This experiment can be understood using the properties of transverse waves. A plane monochromatic (single frequency) wave is described by giving the direction of travel, intensity and frequency. To give a *complete* description, we must specify the manner in which the electric field oscillates in the transverse plane. One simple case is when the tip of the electric vector oscillates back and forth in a straight line, namely, the x -axis. We call this wave as *linearly polarised* along x . It is denoted by a double headed arrow since the electric vector points along x and $-x$ at different times in a cycle. We can similarly think of a wave linearly polarised along the y -direction or any other direction in the transverse plane.

The space variation of the electric field in a linearly polarised wave is shown in Fig. 11.22.

Now, take a wave linearly polarised along a direction making an angle θ with the x -axis. We can resolve this into two waves. One has electric field $E \cos \theta$ and is linearly polarised along x and the other $E \sin \theta$ and is linearly polarised along y . We can now understand the experiment with two polaroids of Fig. 11.21. Let us assume that each polaroid has a special direction in its plane, which we call its *axis*. It is assumed that it transmits only the component of the electric field along the axis and the component perpendicular to the axis is absorbed. In Fig. 11.21(b), the light transmitted by P_2 is polarised along its axis, as shown by the double arrows. The second polaroid P_1 transmits only a fraction $\cos \theta$ of the amplitude, polarised along its axis. The intensity transmitted is a fraction $\cos^2 \theta$. As a function of the angle θ , this is a cosine curve with two maxima (at $\theta = 0^\circ$ and 180°) per rotation. To see this we can write

$$\cos^2 \theta = \frac{1}{2}(1 + \cos 2\theta) \quad (11.24)$$

Notice that for one of the beams, normal incidence gives rise to bending away from the normal. This ray which does not obey Snell's law is called the extraordinary ray. Huygens was able to explain this behaviour by assuming that the wavefront giving rise to this ray was not of a spherical shape.

The two polaroids in an arrangement like Fig. 11.21 are called the polariser and analyser.

The basic reason for the action of the polaroid is that the molecules of a coloured substance (dye) point along a particular axis when the material is prepared. The movement of the electrons along the molecule allows them to absorb the radiation polarised in that direction. As a first approximation, polaroid transmits one component without absorption and absorbs the perpendicular component completely.

Two fairly common devices use polarised light. One is the liquid crystal display (LCD) found in many watches and calculators, and portable (laptop) computers. Liquid crystals have long molecules whose direction can be controlled by applying electric fields. This is used to modify the light produced by a polariser so that its polarisation is perpendicular to the axis of an analyser which cuts it out. These dark regions

can be controlled with applied voltages and used to form letters and numbers.

Some sunglasses (dark glasses) have polaroids. Glare from sunlight reflected from water, snow or some horizontal surface can be reduced if the polaroid cuts out horizontally polarised light (see next section).

We now come to the behaviour of light from a bulb or the Sun passing through a single polaroid (Fig. 11.22). Half the light from the bulb is transmitted for *all* positions of the polaroid. This is because the electric field of the incident light does not have any fixed direction. We have already remarked that the radiation from a source like a bulb is produced by electrons in motion. We expect the direction of the electric field in the transverse plane to change in a more or less irregular way because of the irregular motion of the radiating charges. Light of this kind is called *unpolarised*.

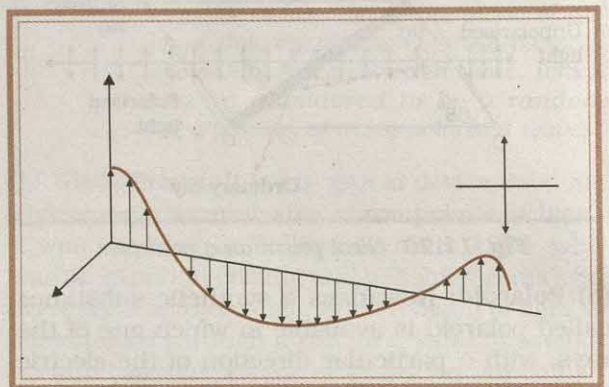


Fig. 11.22 Space variation of the electric field for light which is linearly polarised.

The average transmitted intensity remains constant as we observe unpolarised light through a polaroid which is rotated. The average of Eq. (11.24) for the transmitted intensity, when we allow θ to vary over all angles (0 to 2π) equals $1/2$, since the average of $\cos 2\theta$ is zero.

Example 11.7 If a light beam shows no intensity variation when transmitted through a polaroid which is rotated, does it mean that the light is unpolarised?

Answer No. Consider light which is made up of E_x , E_y with a 90° phase difference but equal amplitudes. The tip of the electric vector executes uniform circular motion at the frequency of the

light itself. This kind of light is called circularly polarised. Because the angle with a fixed polaroid varies rapidly over 0 to 2π radian, the transmitted average intensity is constant and does not change as the polaroid is turned. ◀

11.6.2 Polarisation by Scattering

The light from a clear blue portion of the sky shows a rise and fall of intensity when viewed through a polaroid which is rotated. This is nothing but sunlight, which has changed its direction (having been scattered) on encountering the molecules of the earth's atmosphere. As Fig. 11.23(a) shows, the incident sunlight is unpolarised. The dots stand for polarisation perpendicular to the plane of the figure. The double arrows show polarisation in the plane of the figure. (There is no phase relation between these two in unpolarised light). Under the influence of the electric field of the incident wave the electrons in the molecules acquire components of motion in both these directions. We have drawn an observer looking at 90° to the direction of the Sun. Clearly, charges accelerating parallel to the double arrows do not radiate energy towards this observer since their acceleration has no transverse component. The radiation scattered by the molecule is therefore represented by dots. It is polarised perpendicular to the plane of the figure. This explains the polarisation of scattered light from the sky.

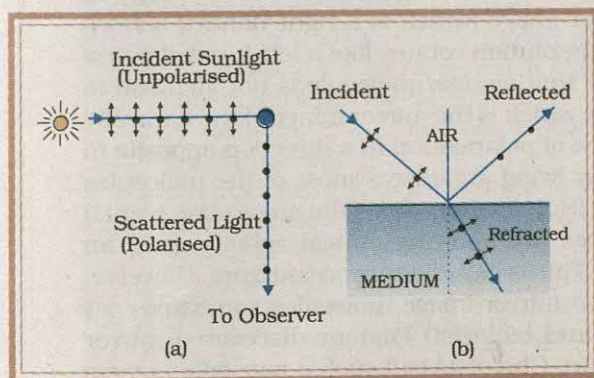


Fig. 11.23 (a) Polarisation of the blue scattered light from the sky. The incident sunlight is unpolarised (dots and arrows). A typical molecule is shown. It scatters light by 90° polarised normal to the plane of the paper (dots only). (b) Polarisation of light reflected from a transparent medium at the Brewster angle (reflected ray perpendicular to refracted ray).

The scattering of light by molecules was intensively investigated by C.V. Raman and his collaborators in Calcutta in the nineteen twenties. Raman was awarded the Nobel Prize for Physics in 1930 for this work.

11.6.3 Polarisation by Reflection

These phenomena are summarised in Brewster's Law (named after the discoverer).

Figure 11.23(b) shows light reflected from a transparent medium, say, water. As before, the dots and arrows indicate that both polarisations are present in the incident and refracted waves. We have drawn a situation in which the reflected wave travels at right angles to the refracted wave. The oscillating electrons in the water produce the reflected wave. These move in the two directions transverse to the radiation from wave in the medium, i.e., the *refracted wave*. The arrows are parallel to the direction of the *reflected wave*. Motion in this direction does not contribute to the reflected wave. As the figure shows, the reflected light is therefore linearly polarised perpendicular to the plane of the figure (represented by dots). This can be checked by looking at the reflected light through an analyser. The transmitted intensity will be zero when the axis of the analyser is in the plane of the figure, i.e., the plane of incidence.

Example 11.8 What should be the angle of incidence in Fig. 11.23(b) so that the reflected and refracted rays are perpendicular?

Answer From the geometry of the Fig. 11.23(b), $i + r = 90^\circ$

$$n = \frac{\sin i}{\sin r} = \frac{\sin i}{\sin (90^\circ - i)} = \tan i$$

In Example 11.8, the special angle of incidence satisfying the condition, $n = \tan i$, is called the *Brewster angle*. For $n = 1.5$, it is approximately 57° .

These results are summarised in Brewster's Law (Example 11.8), named after the English scientist who discovered and studied this phenomenon of polarisation by reflection. ◀

When unpolarised light is incident on the boundary between two transparent media, the reflected light is polarised with its electric vector perpendicular to the plane of incidence when

the refracted and reflected rays make a right angle with each other.

For simplicity, we have discussed scattering of light by 90° , and reflection at the Brewster angle. In this special situation, one of the two perpendicular components of the electric field is zero. At other angles, both components are present but one is stronger than the other. There

is no stable phase relationship between the two perpendicular components since these are derived from two perpendicular components of an unpolarised beam. When such light is viewed through a rotating analyser, one sees a maximum and a minimum of intensity but not complete darkness. This kind of light is called partially polarised.

ROTATION OF THE PLANE OF POLARISATION

This topic is covered extensively in chemistry texts, so we give a brief account here just for completeness. When linearly polarised light travels through glass or water, the direction of linear polarisation remains fixed. However, some substances like sugar solution show the remarkable property of rotating this direction in the transverse plane. This effect is called *optical rotation* or *optical activity*. Fig. 11.24 illustrates an experiment to measure this rotation. A polariser produces linearly polarised light which travels along a tube of sugar solution. The polarisation of the emerging beam can be found using an analyser. The angle of rotation is found to be proportional to the length traversed and to the concentration of sugar.

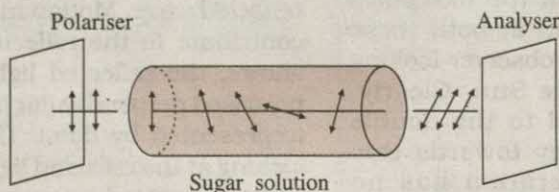


Fig. 11.24 Linearly polarised light from the first polaroid enters sugar solution in a cylindrical tube. The direction of polarisation rotates. This is detected by a second polaroid which gives maximum intensity only after turning it through an angle.

The pattern of polarisation vectors in Fig. 11.24 is like a left-handed screw. As you know, such an object is not identical to its mirror image (which is a right-handed screw). One can ask why the polarisation of light in sugar solution rotates like a left-handed screw rather than a right-handed screw. The answer is that a sugar molecule is not identical to its mirror image. Chemists can prepare a molecule which is the mirror image of the naturally occurring sugar. In solution, this rotates the plane of polarisation in a direction opposite to that shown in Fig. 11.24. It is interesting that in living organisms most of the molecules occur in only one of the two mirror image forms. (Sugar comes from the sugarcane plant!) A chemical preparation starting from substances not showing optical activity gives an equal mixture of the two mirror image molecules. This is called a *racemic* mixture. However, when crystals form from such a mixture, the two mirror image molecules sometimes get separated. This was how the French chemist (and biologist) Pasteur discovered mirror image molecules and their opposite optical rotation. Chemists call such a pair *enantiomers* and the two forms are called *dextro* and *laevo* (which mean right and left, respectively, and are abbreviated by *d* and *l*) depending on the sign of optical activity.

A very simple molecule like nitrogen, N_2 , consists of two atoms joined by a chemical bond. It looks identical to its mirror image. In a gas or in solution, these molecules occur in all possible orientations. This entire assembly looks identical to its mirror image. It cannot produce optical rotation. Only molecules which are not identical to their mirror images can produce this effect. This effect is also shown by crystals like quartz, which again exist in two different forms, which are mirror images of each other.

SUMMARY

1. Huygens' principle tells us that each point on a wavefront is a source of secondary waves, which add up to give the wavefront at a later time.
2. Huygens' construction tells us that the new wavefront is the forward envelope of the secondary waves. When the speed of light is independent of direction, the secondary waves are spherical. The rays are then perpendicular to both the wavefronts and the time of travel is the same measured along any ray. This principle leads to the well known laws of reflection and refraction.
3. The principle of superposition of electric fields applies whenever two or more sources of light illuminate the same point. When we consider the intensity of light due to these sources at the given point, there is an interference term in addition to the sum of the individual intensities. But this term is important only if it has a non-zero average, which occurs only if the sources have the same frequency and a stable phase difference.
4. Young's double slit of separation ' d ' gives equally spaced fringes of angular separation $\frac{\lambda}{d}$. The source, mid-point of the slits, and central bright fringe lie in a straight line. An extended source will destroy the fringes if it subtends angle more than $\frac{\lambda}{d}$ at the slits. A non-monochromatic source with a range of wavelengths $\Delta\lambda$ around λ will give approximately $\frac{\lambda}{\Delta\lambda}$ fringes. For higher values of path differences the fringes are lost.
5. A single slit of width a gives a diffraction pattern with a central maximum. The intensity falls to zero at angles of $\pm\frac{\lambda}{a}, \pm\frac{2\lambda}{a}$ etc., with successively weaker secondary maxima in between. Diffraction limits the angular resolution of a telescope to $\frac{\lambda}{D}$ where D is the diameter. Two stars closer than this give strongly overlapping images. Similarly, a microscope objective subtending 2β at the focus, in a medium of refractive index n , will just separate two objects spaced at a distance $\lambda/2n \sin \beta$, which is the resolution limit of a microscope. Diffraction determines the limitations of the concept of light rays. A beam of width a travels a distance a^2/λ , called the Fresnel distance, before it starts to spread out due to diffraction.
6. Natural light, e.g., from the Sun is unpolarised. This means the electric vector takes all possible directions in the transverse plane, rapidly and randomly, during a measurement. A Polaroid or Nicol prism transmits only one component (parallel to a special axis). The resulting light is called linearly polarised or plane polarised. When this kind of light is viewed through a second Polaroid or Nicol whose axis turns through 360° , two maxima and minima of intensity are seen. Polarised light can also be produced by reflection at a special angle (called the Brewster angle) and by scattering through 90° in the earth's atmosphere.

POINTS TO PONDER

1. Waves from a point source spread out in all directions, while light was seen to travel along narrow rays. It required the insight and experiment of Huygens, Young and Fresnel to understand how a wave theory could explain all aspects of the behaviour of light.

2. The crucial new feature of waves is interference of amplitudes from different sources which can be both constructive and destructive, as shown in Young's experiment.
3. Even a wave falling on single slit should be regarded as a large number of sources which interfere constructively in the forward direction ($\theta = 0$), and destructively in other directions.
4. Diffraction phenomena define the limits of ray optics. The limit of the ability of microscopes and telescopes to distinguish very close objects is set by the wavelength of light.
5. Most interference and diffraction effects exist even for longitudinal waves like sound in air. But polarisation phenomena are special to transverse waves like light wave.

EXERCISES

- 11.1** What is the geometrical shape of the wavefront in each of the following cases:
- (a) Light diverging from a point source.
 - (b) Light emerging out of a convex lens when a point source is placed at its focus.
 - (c) The portion of the wavefront of light from a distant star intercepted by the Earth.
- 11.2** Light of wavelength 5000 \AA falls on a plane reflecting surface. What are the wavelength and frequency of the reflected light? For what angle of incidence is the reflected ray normal to the incident ray?
- 11.3** (a) The refractive index of glass is 1.5. What is the speed of light in glass? (Speed of light in vacuum is $3.0 \times 10^8 \text{ m s}^{-1}$)
- (b) Is the speed of light in glass independent of the colour of light? If not, which of the two colours red and violet travels slower in a glass prism?
- 11.4** Monochromatic light of wavelength 600 nm is incident from air on a glass surface. What are the wavelength, frequency and speed of the refracted light? Refractive index of glass is 1.5.
- 11.5** A region is illuminated by two sources of light. The intensity I at each point is found to be equal to $I_1 + I_2$, where I_1 is the intensity of light at the point when source 2 is absent. I_2 is similarly defined. Are the sources coherent or incoherent? Explain.
- 11.6** In a Young's double-slit experiment, the slits are separated by 0.28 mm and the screen is placed 1.4 m away. The distance between the central bright fringe and the fourth bright fringe is measured to be 1.2 cm . Determine the wavelength of light used in the experiment.
- 11.7** What is the Brewster angle for air to glass transition? (Refractive index of glass = 1.5.)
- 11.8** Two polaroids are placed 90° to each other and the transmitted intensity is zero. What happens when one more polaroid is placed between these two bisecting the angle between them?
- 11.9** Estimate the distance for which ray optics is good approximation for an aperture of 4 mm and wavelength 400 nm .
- 11.10** Two towers on top of two hills are 40 km apart. The line joining them passes 50 m above a hill halfway between the towers. What is the longest wavelength of radio waves, which can be sent between the towers without appreciable diffraction effects?

- 11.11** The 6563 \AA H_α line emitted by hydrogen in a star is found to be red-shifted by 15 \AA . Estimate the speed with which the star is receding from the Earth.

ADDITIONAL EXERCISES

- 11.12** Monochromatic light of wavelength 589 nm is incident from air on a water surface. What are the wavelength, frequency and speed of (a) reflected, and (b) refracted light? Refractive index of water is 1.33 .
- 11.13** Explain how Newton's corpuscular theory predicts the speed of light in a medium, say, water, to be greater than the speed of light in vacuum. Is the prediction confirmed by experimental determination of the speed of light in water? If not, which alternative picture of light is consistent with experiment?
- 11.14** You have learnt in the text how Huygens' principle leads to the laws of reflection and refraction. Use the same principle to deduce directly that a point object placed in front of a plane mirror produces a virtual image whose distance from the mirror is equal to the object distance from the mirror.
- 11.15** Let us list some of the factors, which could possibly influence the speed of wave propagation:
- nature of the source.
 - direction of propagation.
 - motion of the source and/or observer.
 - wavelength.
 - intensity of the wave.
- On which of these factors, if any, does
- the speed of light in vacuum,
 - the speed of light in a medium (say, glass or water), depend?
- 11.16** For sound waves, the Doppler formula for frequency shift differs slightly between the two situations: (i) source at rest; observer moving, and (ii) source moving; observer at rest. The exact Doppler formulas for the case of light waves in vacuum are, however, strictly identical for these situations. Explain why this should be so. Would you expect the formulas to be strictly identical for the two situations in case of light travelling in a medium?
- 11.17** Answer the following questions:
- When monochromatic light is incident on a surface separating two media, the reflected and refracted light both have the same frequency as the incident frequency. Explain why?
 - When light travels from a rarer to a denser medium, it loses some speed. Does the reduction in speed imply a reduction in the energy carried by the light wave?
 - A narrow pulse of light is sent through a medium. Will you expect the pulse to retain its shape as it travels through the medium?
 - In the wave picture of light, intensity of light is determined by the square of the amplitude of the wave. What determines the intensity of light in the photon picture of light?
 - The speed of light in still water is c/n , where n is the refractive index of the water. What is the speed of light in a stream of water flowing at a steady speed v relative to the observer?
- 11.18** What is the effect on the interference fringes in a Young's double-slit experiment due to each of the following operations:
- the screen is moved away from the plane of the slits;
 - the (monochromatic) source is replaced by another (monochromatic) source of shorter wavelength;

- (c) the separation between the two slits is increased;
 - (d) the source slit is moved closer to the double-slit plane;
 - (e) the width of the source slit is increased;
 - (f) the widths of two slits are increased;
 - (g) the monochromatic source is replaced by source of white light?
- (In each operation, take all parameters, other than the one specified, to remain unchanged.)

11.19 Figure 11.25 shows an outline of Lloyd's mirror experiment. M is a plane mirror; S is a narrow slit illuminated by some source of light (not shown) and S' is the image of S in M. M, S and S' are in a plane perpendicular to the paper. O is the line of intersection of the mirror and the screen.

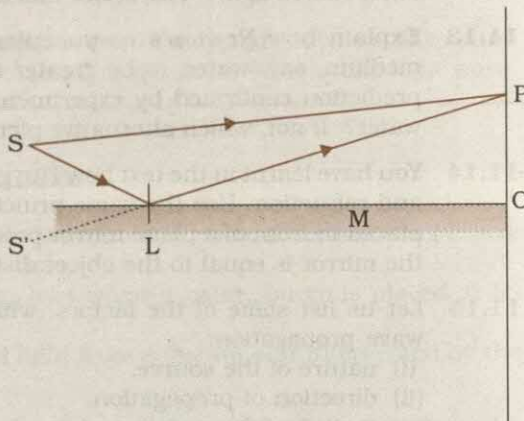


Fig. 11.25

- (a) What is the origin of the fringes observed on the screen?
- (b) Why is the slit S placed so as to have very oblique angle of incidence of light striking the mirrors?
- (c) The two path lengths PS and PS' are equal when P coincides with O. Yet the fringe at O is found in the experiment to be dark, not bright. What does this observation imply?

11.20 Figure 11.26 shows two flat glass plates P_1 and P_2 placed nearly (but not exactly) parallel forming an air wedge. The plates are illuminated normally by monochromatic light and viewed from above. Light waves reflected from the upper and lower surfaces of the air wedge give rise to an interference pattern:

- (a) Show that the separation between two successive bright (or dark) fringes is given by $\frac{\lambda l}{2s}$ where l is the length of each plate and s is the separation between the plates at the open end of the wedge.
- (b) In the experiment, a dark fringe is observed along the line joining the two plates. Why?
- (c) If the space between the glass plates is filled with water, what changes in the fringe pattern do you expect to see, if at all?
- (d) Suggest a way of obtaining a bright fringe along the line of contact of the two plates in this experiment.

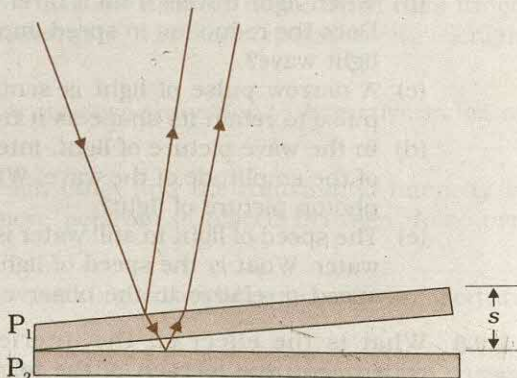


Fig. 11.26

11.21 Give the shape of interference fringes observed

- (a) in a Young's double-slit experiment.
- (b) in the air wedge experiment.
- (c) in the Lloyd's mirror experiment.
- (d) when a small lamp is placed before a thin mica sheet and light waves reflected from the front and back surfaces of the sheet combine to produce interference pattern on a screen behind the lamp. (Pohl's experiment.)
- (e) from a thin air film formed by placing a convex lens on top of a flat glass plate (Newton's arrangement).

11.22 (a) Red light of wavelength 6500 \AA from a distant source falls on a slit 0.50 mm wide. What is the distance between the two dark bands on each side of the central bright band of the diffraction pattern observed on a screen placed 1.8 m from the slit?

- (b) What is the answer to (a) if the slit is replaced by a small circular hole of diameter 0.50 mm ?

11.23 Answer the following questions:

- (a) In a single-slit diffraction experiment, the width of the slit is made double the original width. How does this affect the size and intensity of the central diffraction band?
- (b) In what way is diffraction from each slit related to the interference pattern in a double-slit experiment?
- (c) When a tiny circular obstacle is placed in the path of light from a distant source, a bright spot is seen at the centre of the shadow of the obstacle. Explain why?
- (d) Two students are separated by a 7 m partition wall in a room 10 m high. If both light and sound waves can bend around obstacles, how is it that the students are unable to see each other even though they can converse easily.
- (e) Ray optics is based on the assumption that light travels in a straight line. Diffraction effects (observed when light propagates through small apertures/slits or around small obstacles) disprove this assumption. Yet the ray optics assumption is so commonly used in understanding location and several other properties of images in optical instruments. What is the justification?

11.24 Answer the following questions:

- (a) When a low flying aircraft passes overhead, we sometimes notice a slight shaking of the picture on our TV screen. Suggest a possible explanation.
- (b) Thin films such as a soap bubble or a thin layer of oil on water show beautiful colours when illuminated by white light. Explain the observation.
- (c) In a thin film interference experiment (the experiment on Newton's rings, for example), the central fringe of the pattern is dark when viewed by reflected light, and bright when viewed by transmitted light. Why?
- (d) If white light is used in the air wedge interference experiment (Exercise 11.20) or the Newton's rings experiment, the colour observed in the reflected light is complementary to that observed in the light transmitted through the same point. Why?
- (e) As you have learnt in the text, the principle of linear superposition of wave displacement is basic to understanding intensity distributions in diffraction and interference patterns. What is the justification of this principle?

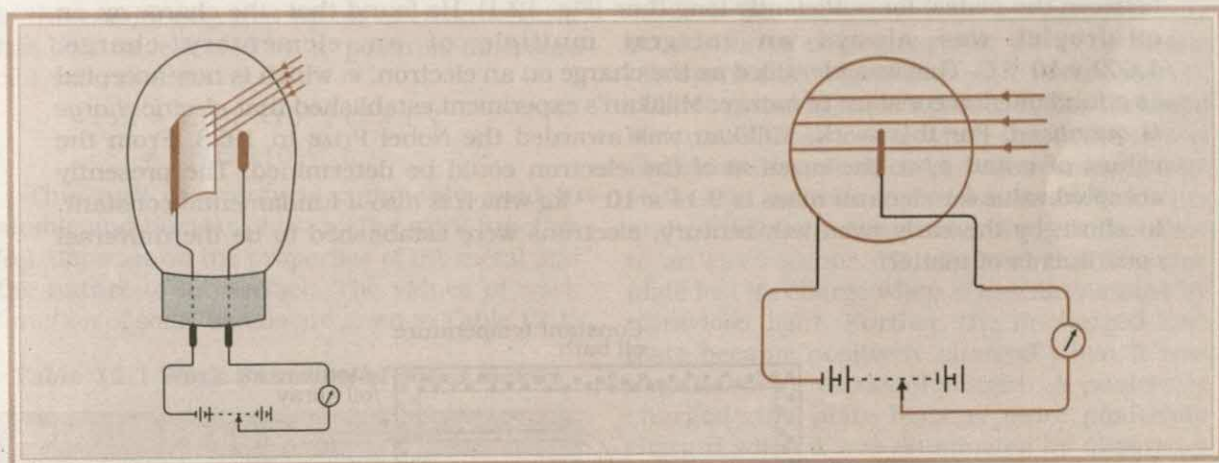
11.25 At a given point in space, circularly polarised light produces equal amplitude vibrations along x and y with a 90° phase difference. $E_x = E_0 \cos \omega t$ and $E_y = E_0 \sin \omega t$. Let x' and y' be a new set of axes rotated by θ in the $x-y$ plane. If the same vibrations $E_0 \cos \omega t$ and $E_0 \sin \omega t$ are present along x' and y' , show that the result is still circularly polarised light with a different phase.

Show that if E_y is changed in phase by π radian, the circle is traversed in the opposite sense.

- 11.26** Show that the two oppositely circularly polarised beams of the same frequency and equal amplitude combine to give linear polarisation. What should one do to the relative phase of the two beams to rotate the direction of linear polarisation? Can you use this to understand what happens to the two opposite circular polarisation in sugar solution?
- 11.27** Two polaroids are placed at 90° to each other and the transmitted intensity is zero.
 (a) What happens when one more polaroid is placed between these two bisecting the angle between them?
 (b) $(N-1)$ more polaroids are inserted between two crossed polaroids (at 90° to each other). Their axes are equally spaced. How does the transmitted intensity behave for large N ? [Hint: Calculate a few special cases, e.g., $N = 4, 8, \dots$]
- 11.28** A half wave plate is a device which introduces a phase difference of π between E_x and E_y . What is its effect on
 (a) linearly polarised light making angle θ to the x -axis?
 (b) circularly polarised light?
- 11.29** Sodium light has two wavelengths $\lambda_1 = 589$ nm and $\lambda_2 = 589.6$ nm. As the path difference increases, when is the visibility of the fringes a minimum?
- 11.30** In deriving the single slit diffraction pattern, it was stated that the intensity is zero at angles of $n\lambda/a$. Justify this by suitably dividing the slit to bring out the cancellation.
- 11.31** In a pinhole camera, a box of length L has a hole of a radius a in one wall. When the hole is illuminated by a parallel beam, the size of spot of light is large. Show that it is also very large when a is small due to diffraction. Assume that the spread due to diffraction just adds to the geometrical spread and find the minimum size of the spot.
- 11.32** Two coherent beams intersect at a small angle θ . What is the spacing of the interference fringes on a screen whose normal bisects the directions of two beams? Instead of a screen, a photographic film is used. When it is developed, the fringes appear as opaque and transparent regions. The film is then used as a grating (a device which consists of a large number of equally spaced single slits). What happens when one of the two beams which produced the interference is allowed to fall on this grating?

CHAPTER TWELVE

DUAL NATURE OF RADIATION AND MATTER



12.1 INTRODUCTION

Experimental investigations on conduction of electricity (electric discharge) through gases at low pressure in a discharge tube led to many historic discoveries towards the end of the nineteenth century. The discovery of X-rays by Roentgen in 1895, and that of electron by J.J. Thomson in 1897, were important milestones in the understanding of atomic structure. It was found at sufficiently low pressure of about 0.001 mm of mercury that a glow discharge took place between the two electrodes by which the electric field was applied to the gas in the discharge tube. The colour of glow depended on the nature of the glass, being yellowish-green for soda glass. The cause of this fluorescence was attributed to the radiation which appeared to be coming from the cathode. These *cathode rays* were discovered, in 1870, by William Crookes who later, in 1879, suggested that these rays consisted of streams of fast moving negatively charged particles. The British physicist J.J. Thomson (1856-1940) confirmed this hypothesis. By applying mutually perpendicular electric and magnetic fields across the discharge tube, J.J. Thomson was the first to determine experimentally the velocity and the specific charge or charge to mass (e/m) ratio of the cathode ray particles. They were found to travel with velocities ranging from about 0.1 to 0.2 times the speed of light (3×10^8 m/s). The presently accepted value of e/m is 1.76×10^{11} C/kg. Further, the value of e/m was found to be independent of the nature of the material/metal used as the cathode, or the gas introduced in the discharge tube. This observation suggested the universality of the cathode ray particles.

Around the same time, in 1887, it was found that certain metals, when irradiated by ultraviolet light, emitted negatively charged particles having small velocities. Also, certain metals when heated to a high temperature were found to emit negatively charged particles. The value of e/m of these particles was found to be the same as that for cathode ray particles. These observations thus established that all these particles, although produced under different conditions, were identical in nature. J.J. Thomson, in 1897, named these particles as *electrons*, and suggested that they were fundamental universal constituents of matter. For his epoch-making discovery of electron he was awarded the Nobel Prize in 1906.

Thomson's experiment enabled him to determine only the ratio e/m but not e or m for the electron separately. In 1913, the American physicist R.A. Millikan (1868-1953) performed the pioneering oil-drop experiment for the precise measurement of the charge on the electron. By applying suitable electric field across two metal plates, the charged oil-droplets could be caused to rise or fall, or even held stationary in the field of view (space between the plates) for sufficiently long time (Fig. 12.1). He found that the charge on an oil-droplet was always an integral multiple of an elementary charge, 1.602×10^{-19} C. This was identified as the charge on an electron, e , which is now accepted as a fundamental constant of nature. Millikan's experiment established that *electric charge is quantised*. For this work, Millikan was awarded the Nobel Prize in 1923. From the values of e and e/m , the mass m of the electron could be determined. The presently accepted value for electron mass is 9.11×10^{-31} kg which is also a fundamental constant. In short, by the early twentieth century, electrons were established to be the universal constituents of matter.

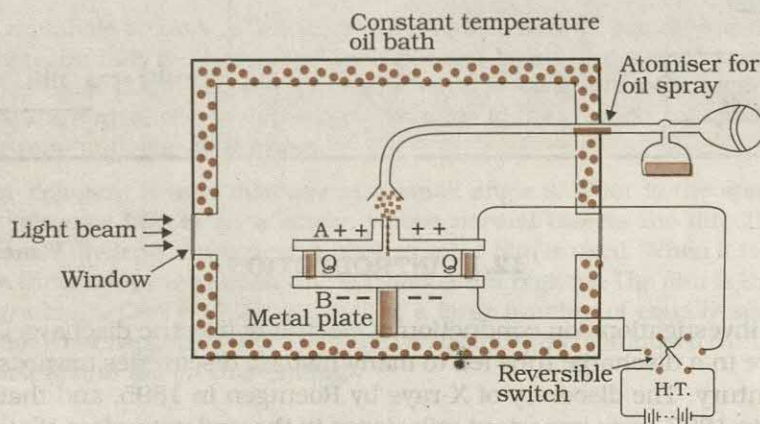


Fig. 12.1 Millikan's oil-drop experiment.



Louis-Victor de Broglie (1892-1987)

French physicist who put forth revolutionary idea of wave nature of matter. This idea was developed by Erwin Schrödinger into a full-fledged theory of quantum mechanics commonly known as wave mechanics. In 1929, he was awarded the Nobel Prize in Physics for his discovery of the wave nature of electrons.

12.2 ELECTRON EMISSION

We know that metals have free electrons that are responsible for their conductivity. However, the free electrons cannot normally escape out of the metal surface. If the electrons come out of the metal, its surface acquires positive charge and attracts the electrons back to the metal. The free electrons are then held inside the metal surface by the attractive forces of the ions. Thus, a certain minimum energy is required to pull the electrons out from within the metal. This minimum energy is called the *work function* of the metal. It is generally denoted by ϕ_0 and measured in eV (electron volt). One electron volt is the energy acquired by an electron when it has been accelerated by a potential difference of 1 volt, so that

$$1 \text{ eV} = 1.602 \times 10^{-19} \text{ C} \times 1 \text{ V} \\ = 1.602 \times 10^{-19} \text{ J}.$$

This unit of energy is commonly used in atomic and nuclear physics. The work function (ϕ_0) depends on the properties of the metal and the nature of its surface. The values of work function of some metals are given in Table 12.1.

Table 12.1 Work Function of some Metals

Metal	Work function ϕ_0 (eV)	Metal	Work function ϕ_0 (eV)
Cs	2.14	Hg	4.49
K	2.30	Cu	4.65
Na	2.75	Ag	4.70
Ca	3.20	Ni	5.15
Pb	4.25	Pt	5.65
Al	4.28		

Note from Table 12.1 that the work function of platinum is the highest ($\phi_0 = 5.65$ eV) while it is the lowest ($\phi_0 = 2.14$ eV) for caesium.

The minimum energy required for the electron emission from the metal surface can be supplied to the free electrons by any one of the following physical processes.

- (i) **Thermionic Emission:** By suitably heating, sufficient thermal energy can be imparted to the free electrons to enable them to come out of the metal.

- (ii) **Field Emission:** By applying a very strong electric field (of the order of 10^8 V m^{-1}) to a metal, electrons can be pulled out of the metal.

- (iii) **Photo-electric Emission:** When light of suitable high frequency shines on the surface of a metal, it emits electrons. These photo(light)-generated electrons are called *photoelectrons*. The phenomenon is called the *photoelectric effect*.

12.3 PHOTOELECTRIC EFFECT

The phenomenon of photoelectric emission was discovered by Heinrich Hertz (1857-1894) in 1887. In his experimental investigation on the production of electromagnetic waves by means of a spark discharge (Chapter 9), Hertz observed that a high voltage spark passed across the metal electrodes more easily when the cathode was illuminated by ultraviolet light from an arc lamp. Hallwachs, in 1888, undertook the study further and connected a negatively charged zinc plate to an electroscope. He observed that the zinc plate lost its charge when it was illuminated by ultraviolet light. Further, the uncharged zinc plate became positively charged when it was irradiated by ultraviolet light. A positively charged zinc plate became more positively charged when it was illuminated by ultraviolet light. From these observations he concluded that negatively charged particles were emitted by the zinc plate under the action of ultraviolet light. After the discovery of electrons by J.J. Thomson, these particles were termed as *photoelectrons*.

The photoelectric effect involves conversion of light energy into electrical energy. It was found that certain metals like zinc, cadmium, magnesium, etc., responded only to ultraviolet light. However, some alkali metals such as lithium, sodium, potassium, caesium and rubidium were sensitive even to the visible light. All these *photosensitive substances* emit electrons when they are illuminated by light.

12.4 EXPERIMENTAL STUDY OF PHOTOELECTRIC EFFECT

Figure 12.2 depicts a schematic view of the arrangement used for the experimental study of the photoelectric effect. It consists of an evacuated glass/quartz tube having a photosensitive plate C and another metal plate A. Monochromatic light radiation of sufficiently

short wavelength passes through the window W and falls on the photosensitive plate C (cathode or emitter). A transparent quartz window is sealed on to the glass tube, which permits ultraviolet radiation to pass through it and irradiate the photosensitive plate C. The electrons are emitted by the cathode C and are collected by the plate A (anode or collector). The potential difference between the plates C and A can be varied. The direction of potential difference can be changed by the commutator. Thus, the plate A can be maintained at a desired positive or negative potential with respect to C. The emission of electrons causes a flow in the outer circuit, establishing an electric current in the circuit. The potential difference between the electrodes is measured by a voltmeter (V) whereas the resulting photocurrent flowing in the circuit is measured by a microammeter (μA). The photoelectric current can be increased or decreased by varying the magnitude and sign of the anode potential with respect to the cathode. It is found that there is a certain minimum negative (retarding) potential sufficient to reduce the current to zero. This is called the *cut-off voltage* or *stopping potential* V_0 .

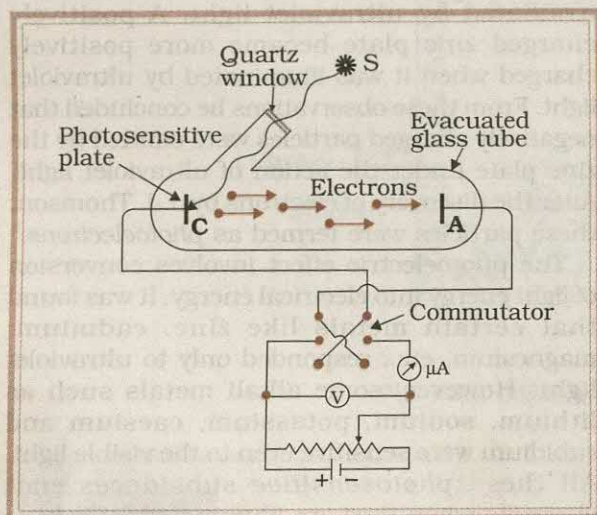


Fig. 12.2 Experimental arrangement for study of photoelectric effect.

Different photosensitive materials respond differently to light. Selenium is more sensitive than zinc or copper. The same photosensitive substance gives different response to light of different wavelengths. For example, ultraviolet light gives rise to photoelectric effect in copper

while green or red light does not. The intensity and frequency of the incident light can be varied, as can the potential difference V between the emitter C and the collector A.

We can use the experimental arrangement of Fig. 12.2 to study the variation of photocurrent with (a) intensity of radiation, (b) frequency of incident radiation, (c) the potential difference between the plates A and C, and (d) the nature of the material of plate C. Light of different wavelengths can be used by putting appropriate filter or coloured glass in the path of light falling on the emitter C. The intensity of light is varied by changing the distance of the light source from the emitter. An extensive study of photoelectric effect was carried out by Lenard and R. A. Millikan.

12.4.1 Effect of Intensity of Light on Photocurrent

The collector A is maintained at a positive potential with respect to emitter C so that electrons ejected from C are attracted towards anode A. Keeping the frequency of the incident radiation and the accelerating potential fixed, the intensity of light is varied and the resulting photoelectric current is measured each time. It is found that the photocurrent increases linearly with intensity of incident light as shown graphically in Fig. 1.2.3. Since photocurrent is directly proportional to the number of photoelectrons emitted per second, this implies that the number of photoelectrons emitted per second is proportional to the intensity of incident radiation.

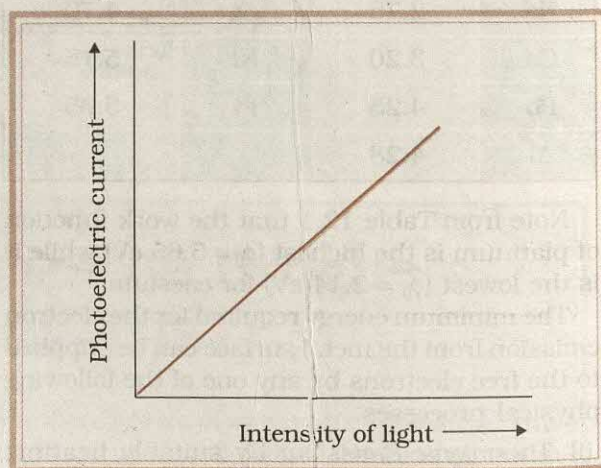


Fig. 12.3 Graph of Photoelectric current versus intensity of light.

12.4.2 Effect of Potential on Photoelectric Current

We first maintain the anode A at some positive accelerating potential with respect to the cathode and illuminate the cathode C with light of fixed frequency ν and fixed intensity I_1 . We next vary the positive potential of anode A gradually and measure the resulting photocurrent each time. It is found that the photoelectric current increases with increase in accelerating (positive) potential until a stage reaches at which, for a certain positive potential of plate A, the photoelectric current becomes maximum or saturates. If we increase the accelerating potential of plate A further, the photocurrent does not increase. This maximum value of the photoelectric current is called the *saturation current*. Saturation current corresponds to the case when all the photoelectrons emitted by the cathode reach the anode.

We now apply a negative (retarding) potential to the anode A with respect to the cathode C and make it increasingly negative gradually. The photocurrent is found to decrease rapidly until it becomes zero at a certain sharply defined negative potential V_0 on the anode A. For a particular frequency of incident radiation, **the minimum negative (retarding) potential V_0 given to the anode A for which the photocurrent becomes zero is called the cut-off or stopping potential.**

The interpretation of the observation in terms of photoelectrons is straightforward. The photoelectrons emitted from the metal do not all have the same energy. Photoelectric current is zero when the stopping potential equals the maximum kinetic energy (K_{\max}) of the photoelectron, so that

$$K_{\max} = e V_0 \quad (12.1)$$

We can now repeat this experiment with incident radiation of the same frequency but of higher intensity I_2 and I_3 ($I_3 > I_2 > I_1$). We note that the saturation currents are now found to be greater (proportional to the intensity of incident radiation) but the stopping potential remains the same as that for the incident radiation of intensity I_1 , as shown graphically in Fig. 12.4. Thus, **for a given frequency of the incident radiation, the stopping potential is independent of its intensity.** In other words,

the maximum kinetic energy of photoelectrons is independent of intensity of incident radiation.

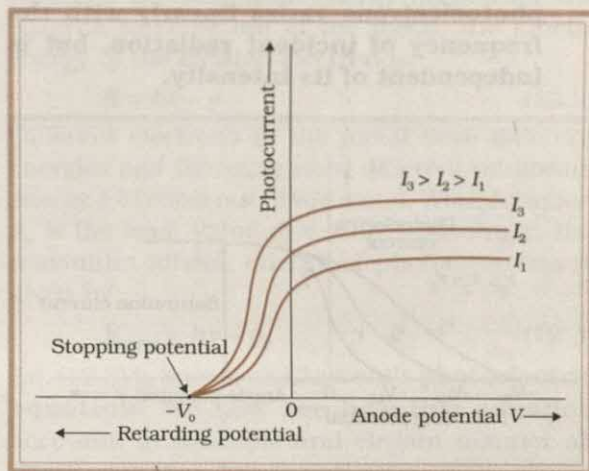


Fig. 12.4 Plot of photocurrent against anode potential for different intensity of incident radiation.

12.4.3 Effect of Frequency of Incident Radiation on Stopping Potential

We now study the relation between the frequency ν of the incident radiation and the stopping potential V_0 . We suitably adjust the same intensity of light radiation at various frequencies and study the variation of photocurrent with anode potential. The resulting variation is shown in Fig. 12.5. We obtain different values of stopping potential but the same value of the saturation current for incident radiation of different frequencies. The stopping potential is more negative for higher frequencies of incident radiation. Note from Fig. 12.5 that the stopping potentials are in the order $V_{03} > V_{02} > V_{01}$ if the frequencies are in the order $\nu_3 > \nu_2 > \nu_1$. This implies that greater the frequency of incident light, greater is the maximum kinetic energy of the photoelectrons. Consequently, we need greater retarding potential to stop them completely. If we plot a graph between the frequency of incident radiation and the corresponding stopping potential for different metals we get a straight line as shown in Fig. 12.6.

The graph shows that

- the stopping potential V_0 varies linearly with the frequency of incident radiation for a given photosensitive material.
- there exists a certain minimum cut-off frequency ν_0 for which the stopping potential is zero.

These observations have two implications:

- (1) **The maximum kinetic energy of the photoelectrons varies linearly with the frequency of incident radiation, but is independent of its intensity.**

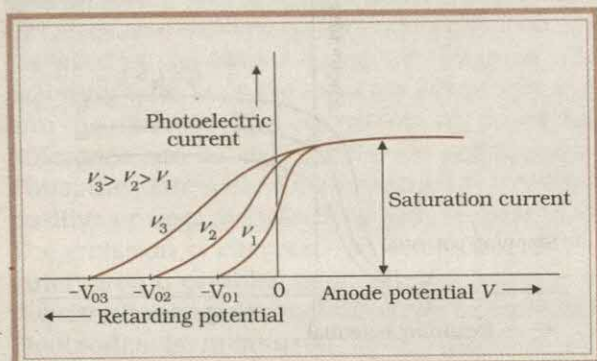


Fig. 12.5 Plot of photoelectric current against anode potential for different frequencies of incident radiation.

- (2) **For a frequency ν of incident radiation, lower than the cut-off frequency ν_0 , no photoelectric emission is possible even if the intensity is large enough.** This minimum, cut-off frequency ν_0 , is called the **threshold frequency** ν_0 . It is different for different metals.

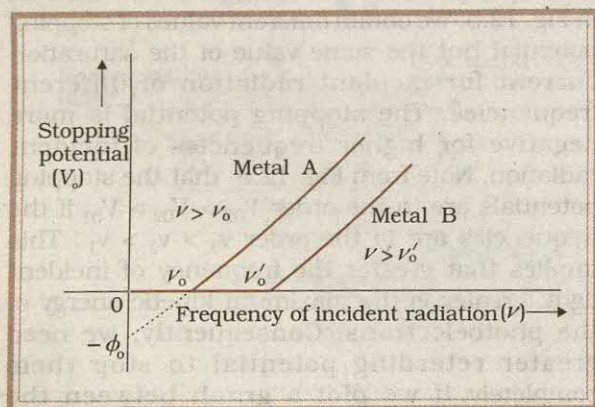


Fig. 12.6 Variation of stopping potential V_0 with frequency ν of incident radiation for a given photosensitive material.

Note that in all the above experiments, it is found that, if frequency of the incident radiation exceeds the threshold frequency, the photoelectric emission starts instantaneously without any apparent time lag, even if the incident radiation is very dim. It is now known that emission starts in a time of the order of 10^{-9} s.

12.4.4 Experimental Features of Photoelectric Emission

We summarise the experimental features and observations described in section 12.4.

- (i) For a given photosensitive material and frequency of incident radiation (above the threshold frequency), the photoelectric current is directly proportional to the intensity of light (Fig. 12.3).
- (ii) For a given photosensitive material and frequency of incident radiation, saturation current is found to be proportional to the intensity of radiation whereas the stopping potential is independent of intensity (Fig. 12.4).
- (iii) For a given photosensitive material, there exists a certain minimum cut-off frequency, called the *threshold frequency*, below which no emission of photoelectrons takes place, no matter how intense the light is. Above the threshold frequency, the stopping potential or equivalently the maximum kinetic energy of the emitted photoelectrons increases linearly with the frequency of the incident radiation, but is independent of its intensity (Fig. 12.6).
- (iv) The photoelectric emission is an instantaneous process without any apparent time lag ($\sim 10^{-9}$ s), even when the incident radiation is made exceedingly dim.

The significance of these experimental results on photoelectric effect becomes evident when we compare them with the predictions of the classical wave theory.

12.5 PHOTOELECTRIC EFFECT AND WAVE THEORY OF LIGHT

The wave nature of light was well established by the end of the nineteenth century. The phenomena of interference, diffraction and polarisation were explained in a natural and satisfactory way by the wave picture of light. According to this picture, light is an electromagnetic wave consisting of electric and magnetic fields with continuous distribution of energy over the region of space over which the wave is extended. Let us now see if this wave picture of light can explain the observations on photoelectric emission given in Section 12.4.

According to the wave picture of photoelectric emission, the free electrons at the surface of the

metal (over which the beam of radiation impinges) absorb the radiant energy continuously. The greater the intensity of radiation, the greater are the amplitude of electric and magnetic fields, and greater is the energy density of the wave. Consequently, the greater the intensity, the greater should be the energy absorbed by each electron. In this picture, the maximum kinetic energy of the photoelectron is then expected to increase with increase in intensity. Also, no matter what the frequency of radiation is, a sufficiently intense beam of radiation (over sufficient time) should be able to impart enough energy to the electrons, so that they exceed the minimum energy needed to get out of the metal. A threshold frequency, therefore, should not exist. These expectations of the wave theory directly contradict observations (i), (ii) and (iii) given in sub-section 12.4.4.

Further, we should note that in the wave picture, the absorption of energy by electron takes place continuously over the entire wave front of the radiation. Since a large number of electrons absorb energy, the energy absorbed per electron per unit time turns out to be small. Explicit calculations estimate that it can take hours or more (such estimates vary depending on the assumption regarding how many electrons absorb the energy) for a single electron to pick up sufficient energy to overcome the work function and come out of the metal. This conclusion is again in striking contrast to observation (iv) that says that the photoelectric emission is (nearly) instantaneous. In short, the wave picture is unable to explain the most basic features of photoelectric emission.

12.6 EINSTEIN'S PHOTOELECTRIC EQUATION: ENERGY QUANTUM OF RADIATION

In 1905, Albert Einstein (1879-1955) proposed a radically new picture of electromagnetic radiation to explain photoelectric effect. In this picture, photoelectric emission does not take place by continuous absorption of energy from radiation. Radiation energy is built up of discrete units – the so called **quanta of energy of radiation**. Each quantum of radiant energy has energy $h\nu$, where h is Planck's constant and ν the frequency of light. In photoelectric effect, an electron absorbs a quantum of energy ($h\nu$)

of radiation. If this quantum of energy absorbed exceeds the minimum energy needed for the electron to come out of the metal, the kinetic energy of the emitted electron is:

$$K = h\nu - \phi \quad (12.2)$$

Different electrons in the metal have different energies and therefore need different minimum energy ϕ to come out of the metal. Work function ϕ_0 is the least value of ϕ . Correspondingly, the maximum kinetic energy of photoelectrons is given by

$$K_{\max} = h\nu - \phi_0 \quad (12.3)$$

Eq. (12.3) is known as **Einstein's photoelectric equation**. We now see how this equation accounts in a simple and elegant manner all the observations on photoelectric effect given in sub-section 12.4.4.

- According to Eq. (12.3), K_{\max} depends linearly on ν , and is independent of intensity of radiation, in agreement with observation. This has happened because in Einstein's picture, photoelectric effect arises from the absorption of a single quantum of radiation by a single electron. The intensity of radiation (that is proportional to the number of energy quanta per unit area per unit time) is irrelevant to this basic process.
- Since K_{\max} must be non-negative, Eq. (12.3) implies that photoelectric emission is possible only if

$$h\nu > \phi_0$$

or $\nu > \nu_0$, where

$$\nu_0 = \frac{\phi_0}{h} \quad (12.4)$$

Thus, there exists a threshold frequency ν_0 ($= \phi_0/h$) below which no photoelectric emission is possible, whatever the intensity of radiation may be.

- In this picture, intensity of radiation as noted above, is proportional to the number of energy quanta per unit area per unit time. The greater the number of energy quanta available, the greater is the number of electrons absorbing the energy quanta and greater, therefore, is the number of electrons coming out of the metal (for $\nu > \nu_0$). This explains why, for $\nu > \nu_0$, photoelectric current is proportional to intensity.

- In Einstein's picture, the basic elementary process involved in photoelectric effect is the absorption of a light quantum by an electron. This process is instantaneous. Thus, whatever may be the intensity i.e., the number of quanta of radiation per unit area per unit time, photoelectric emission is instantaneous. Low intensity does not mean delay in emission, since the basic elementary process is the same. Intensity only determines how many electrons are able to participate in the elementary process (absorption of a light quantum by a single electron) and, therefore, the photoelectric current.

Using Eq. (12.1), the photoelectric equation, can be written as

$$eV_0 = h\nu - \phi_0$$

$$\text{or } V_0 = \left(\frac{h}{e}\right)\nu - \frac{\phi_0}{e} \quad (12.5)$$

This is an important result. It predicts that the V_0 versus ν curve is a straight line with

slope = (h/e) , and intercept = $-\frac{\phi_0}{e}$. In 1916,

R.A. Millikan measured the slope of the straight line obtained for sodium. Using the known value of e , he determined the value of Planck's constant h . This value was close to the value of Planck's constant ($= 6.626 \times 10^{-34} \text{ J s}$) determined in an entirely different context. In the same way,

from the intercept ($= -\frac{\phi_0}{e}$) and the known value of e , he obtained the value of the work function ϕ_0 for sodium, that agreed well with the known value.

The successful explanation of photoelectric effect using the hypothesis of light quanta and the experimental determination of values of h and ϕ_0 , in agreement with values obtained from other experiments, led to the acceptance of Einstein's picture of photoelectric effect. Millikan verified photoelectric equation with great precision, for a number of alkali metals over a wide range of radiation frequencies. He was awarded the Nobel Prize in Physics in 1923 for his work on the determination of e and photoelectric effect.

12.7 THE PHOTON

Photoelectric effect thus gave evidence to the strange fact that light in interaction with matter behaved as if it was made of quanta or packets of energy, each of energy $h\nu$.

Is the light quantum of energy to be associated with a particle? Einstein was not fully convinced about it at this stage. Some years later, in another theoretical investigation that we cannot describe here, Einstein arrived at the important result, that the light quantum can also be associated with momentum ($h\nu/c$). A definite value of energy as well as momentum is a strong sign that the light quantum can be associated with a particle. This particle was later named 'photon'. The particle-like behaviour of light was further confirmed in 1924 by the experiment of A. H. Compton (1892-1962) on scattering of X-rays from electrons. For his work on photoelectric effect, Einstein received the Nobel Prize in Physics in 1921.

We can summarise the photon picture of electromagnetic radiation as follows:

- In interaction of radiation with matter, radiation behaves as if it is made up of particles called photons.
- Each photon has energy $E (=h\nu)$ and momentum $p (=h\nu/c)$, and speed c , the speed of light. From Einstein's relativistic energy-momentum relation, $E = \sqrt{c^2 p^2 + m_0 c^4}$, formally the rest mass m_0 of the photon is zero. However, the rest mass m_0 of photon is not a physically meaningful quantity, since we can never have a frame of reference in which photon is at rest. It moves with speed c relative to all frames.
- The moving mass m of photon is $m = E/c^2 = h\nu/c^2$
- All photons of light of a particular frequency ν , have the same energy $E (=h\nu)$ and momentum $p (=h\nu/c)$, whatever the intensity of radiation may be.
- Photons are electrically neutral and are not deflected by electric and magnetic fields.
- In a photon-particle collision (such as photon-electron collision), the total energy and total momentum are conserved. However, the number of photons may not be conserved in a collision. The photon may be absorbed or a new photon may be created.

Example 12.1 Monochromatic light of frequency 6.0×10^{14} Hz is produced by a laser. The power emitted is 2.0×10^{-3} W. (a) What is the energy of a photon in the light beam? (b) How many photons per second, on the average, are emitted by the source?

Answer

- (a) Each photon has an energy

$$E = h\nu = (6.63 \times 10^{-34} \text{ J s}) (6.0 \times 10^{14} \text{ Hz})$$

$$E = 3.98 \times 10^{-19} \text{ J}$$

- (b) If N is the number of photons emitted by the source per second, the power P transmitted in the beam equals N times the energy per photon E , so that $P = NE$. Then

$$N = \frac{P}{E} = \frac{2.0 \times 10^{-3} \text{ W}}{3.98 \times 10^{-19} \text{ J}}$$

$$N = 5.0 \times 10^{15} \text{ photons per second.}$$

Example 12.2 The work function of caesium is 2.14 eV. Find (a) the threshold frequency for caesium, and (b) the wavelength of the incident light if the photocurrent is brought to zero by a stopping potential of 0.60 V.

Answer

- (a) For the minimum, cut-off or threshold frequency, the energy $h\nu_0$ of the incident radiation must be equal to work function ϕ_0 , so that

$$\nu_0 = \frac{\phi_0}{h} = \frac{2.14 \text{ eV}}{6.63 \times 10^{-34} \text{ J s}}$$

$$\nu_0 = \frac{2.14 \times 1.6 \times 10^{-19} \text{ J}}{6.63 \times 10^{-34} \text{ J s}} = 5.16 \times 10^{14} \text{ Hz}$$

Thus, for frequencies less than the threshold frequency $\nu_0 = 5.16 \times 10^{14}$ Hz, no photoelectrons are ejected.

- (b) Photocurrent reduces to zero, when maximum kinetic energy of the emitted photoelectrons equals the potential energy eV_0 by the retarding potential V_0 . Einstein's Photoelectric equation is

$$eV_0 = h\nu - \phi_0 = \frac{hc}{\lambda} - \phi_0$$

$$\text{or, } \lambda = hc/(eV_0 + \phi_0)$$

$$= \frac{(6.63 \times 10^{-34} \text{ J s}) (3 \times 10^8 \text{ m/s})}{(0.60 \text{ eV} + 2.14 \text{ eV})}$$

$$= \frac{19.89 \times 10^{-26} \text{ J m}}{(2.74 \text{ eV})}$$

$$\lambda = \frac{19.89 \times 10^{-26} \text{ J m}}{2.74 \times 1.6 \times 10^{-19} \text{ J}} = 454 \text{ nm}$$

12.8 PHOTO-CELL

A photo-cell is a technological application of the photoelectric effect. It is a device which converts light energy into electrical energy. It is also sometimes called an electric eye. A photo-cell consists of a semi-cylindrical photo-sensitive metal plate C (emitter) and a wire loop A (collector) supported in an evacuated glass or quartz bulb. It is connected to the external circuit having a high-tension battery B and microammeter (μA) as shown in Fig. 12.7(a). Sometimes, instead of the plate C, a thin layer of photosensitive material (C) is pasted on the inside of the bulb. A part of the bulb is left clean for the light to enter it as shown in Fig. 12.7(b).

When light of suitable wavelength falls on the cathode, photoelectrons are emitted. These photoelectrons are drawn to the collector by an electric field. The resulting photocurrent is measured by a sensitive microammeter (μA) in the external circuit. The photocurrent of the order of a few microampere can be normally obtained from a photo-cell.

A photo-cell converts a change in intensity of illumination into a change in photocurrent. This current can be used to operate control systems and in light measuring devices. Light meters in photographic cameras make use of photo-cells to measure the intensity of light. The photo-cells, inserted in the street light electric circuit, are used to switch on and off the street lighting system automatically at dusk and dawn. They are used in the control of a counting device which records every interruption of the light beam caused by a person or object passing across the beam. So photo-cells help count the persons entering an auditorium, provided they enter the hall one by one. They are used for detection of traffic law defaulters: an alarm may be sounded whenever a beam of (invisible) radiation is intercepted.

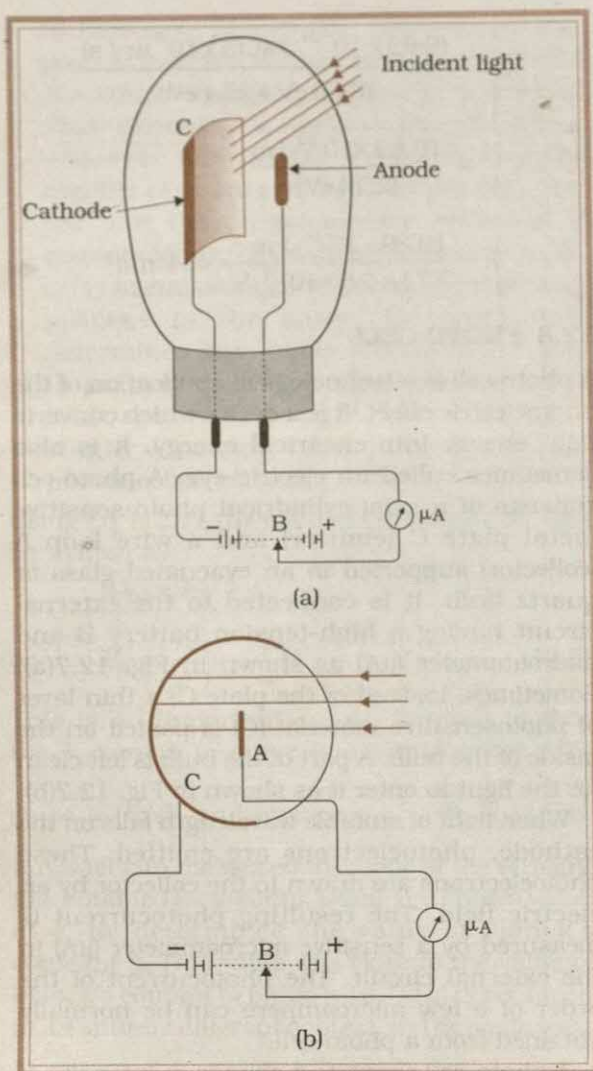


Fig. 12.7 A photo-cell.

In burglar alarm, (invisible) ultraviolet light is continuously made to fall on the photo-cell installed at the door-way. A person entering the door interrupts the beam falling on the photo-cell. The abrupt change in photocurrent is used to start an electric bell ringing. In fire alarm, a number of photo-cells are installed at suitable places in a building. In the event of breaking out of fire, light radiations fall upon the photo-cell. This completes the electric circuit through an electric bell or a siren which starts operating as a warning signal.

The photo-cells are used in the reproduction of sound in cinematography (motion pictures) and in the television camera for scanning and telecasting scenes. They are used in industries

for detecting minor flaws or holes in metal sheets.

12.9 THE WAVE NATURE OF MATTER

The dual (wave-particle) nature of light (electromagnetic radiation, in general) comes out clearly from what we have learnt in this and the preceding Chapters. The wave nature of light shows up in the phenomena of interference, diffraction and polarisation. On the other hand, in photoelectric effect and Compton effect which involve energy and momentum transfer, radiation behaves as if it is made up of a bunch of particles – the photons. Whether a particle or wave description is best suited for understanding an experiment depends on the nature of the experiment. For example, in the familiar phenomenon of seeing an object by our eye, both descriptions are important. The gathering and focussing mechanism of light by the eye-lens is well described in the wave picture. But its absorption by the rods and cones (of the retina) requires the photon picture of light. So far, no single experiment has been devised which displays both wave and particle aspects of radiation at the same time.

A natural question arises: If radiation has a dual (wave-particle) nature, might not the particles of nature (the electrons, protons, etc.) also exhibit wave-like character? In 1924, the French physicist Louis Victor de Broglie (pronounced as de Broij) (1892-1987) put forward the bold hypothesis that moving particles of matter should display wave-like properties under suitable conditions. He reasoned that nature was symmetrical and that the two basic physical entities — matter and energy, must have symmetrical character. If radiation shows dual aspects, so should matter. De Broglie proposed that the wave length λ associated with a particle of momentum p is given as

$$\lambda = \frac{h}{p} = \frac{h}{mv} \quad (12.6)$$

where m is the mass of the particle and v its speed. Eq. (12.6) is known as the **de Broglie relation** and the wavelength λ of the **matter wave** is called **de Broglie wave length**. The dual aspect of matter is evident in the de Broglie relation. On the left hand side of Eq. (12.6), λ is the attribute of a wave while on the right hand side the momentum p is a typical attribute of

particle. The Planck's constant h relates the two attributes.

Equation (12.6) for a material particle is basically a hypothesis whose validity can be tested only by experiment. However, it is interesting to see that it is satisfied also by a photon. For a photon, as we have seen,

$$p = h\nu / c \quad (12.7)$$

Therefore,

$$\frac{h}{p} = \frac{c}{\nu} = \lambda \quad (12.8)$$

That is, the de Broglie wavelength of a photon given by Eq. (12.6) equals the wavelength of electromagnetic radiation of which the photon is a quantum of energy and momentum.

Clearly, from Eq. (12.6), λ is smaller for a heavier particle (large m) or more energetic particle (large ν). For example, the de Broglie wavelength of a ball of mass 0.12 kg moving with a speed of 20 m s^{-1} is easily calculated.

$$p = mv = 0.12 \text{ kg} \times 20 \text{ m s}^{-1} = 2.40 \text{ kg m s}^{-1}$$

$$\lambda = \frac{h}{p} = \frac{6.63 \times 10^{-34} \text{ J s}}{2.40 \text{ kg m s}^{-1}} = 2.76 \times 10^{-34} \text{ m}$$

This wavelength is so small that it is beyond any measurement. This is the reason why macroscopic objects in our daily life do not show wave-like properties. On the other hand, in the sub-atomic domain, the wave character of particles is significant and measurable.

Consider an electron (mass m , charge e) accelerated from rest through a potential V . The kinetic energy K of the electron equals the work done (Ve) on it by the electric field:

$$K = Ve \quad (12.9)$$

$$\text{Now } K = \frac{1}{2}mv^2 = \frac{p^2}{2m}, \text{ so that}$$

$$p = \sqrt{2mK} = \sqrt{2mVe} \quad (12.10)$$

The de Broglie wavelength λ of the electron is then

$$\lambda = \frac{h}{p} = \frac{h}{\sqrt{2mK}} = \frac{h}{\sqrt{2mVe}} \quad (12.11)$$

Substituting the numerical values of h , m , e , we get

$$\lambda = \frac{1.227}{\sqrt{V}} \text{ nm} \quad (12.12)$$

where V is the magnitude of accelerating potential in volts. For a 120 V accelerating potential, Eq. (12.12) gives $\lambda = 0.112 \text{ nm}$. This wavelength is of the same order as the spacing between the atomic planes in crystals. This suggests that matter waves associated with an electron could be verified by crystal diffraction experiments analogous to X-ray diffraction. We describe the experimental verification of the de Broglie hypothesis in the next section. For his discovery of the wave nature of electrons, de Broglie was awarded the Nobel Prize in Physics in 1929.

It is worth pausing here to reflect on just what a matter wave associated with a particle, say, an electron, means. Actually, a truly satisfactory physical understanding of the dual nature of matter and radiation has not emerged so far. The great founders of quantum mechanics (Niels Bohr, Albert Einstein, and many others) struggled with this and related concepts for long. Still the deep physical interpretation of quantum mechanics continues to be an area of active research. Despite this, the concept of matter wave has been mathematically introduced in modern quantum mechanics with great success. An important milestone in this connection was when Max Born (1882-1970) suggested a probability interpretation to the matter wave amplitude. According to this, the intensity (square of the amplitude) of the matter wave at a point determines the probability density* of the particle at that point. Thus, if the intensity of matter wave is large in a certain region, there is a greater probability of the particle being found there than where the intensity is small.

The matter wave picture elegantly incorporated the Heisenberg's Uncertainty Principle that you will learn in detail in more advanced courses. According to the principle, it is not possible to measure **both** the position and momentum of an electron (or any other particle) **at the same time** exactly. There is always some uncertainty (Δx) in the specification of position and some uncertainty (Δp) in the specification of momentum. The product of Δx and Δp has a lower bound, of the order of h^{**} (with $h = h/2\pi$)

*. Probability density means probability per unit volume. Thus, if A is the amplitude of the wave at a point, $|A|^2 \Delta V$ is the probability of the particle being found in a small volume ΔV around that point.

** A more rigorous treatment gives $\Delta x \Delta p \geq h/2$.

i.e.,

$$\Delta x \Delta p = \hbar \quad (12.13)$$

Equation (12.13) allows the possibility that Δx is zero; but then Δp must be infinite in order that the product is non-zero. Similarly, if Δp is zero, Δx must be infinite. Ordinarily, both Δx and Δp are non-zero such that their product is of the order of \hbar .

Now, if an electron has a definite momentum p ($\Delta p = 0$), by the de Broglie relation, it has a definite wavelength λ . A wave of definite (single) wavelength extends all over space. By Born's probability interpretation this means that the electron is not localised in any finite region of space. That is, its position uncertainty is infinite ($\Delta x \rightarrow \infty$), which is consistent with the Uncertainty Principle.

In general, the matter wave associated with the electron is not extended all over space. It is a wave packet extending over some finite region of space. In that case Δx is not infinite but has some finite value depending on the extension of the wave packet. Also, you must appreciate that a wave packet of finite extension does not have a single wavelength. It is built up of wavelengths spread around some central wavelength.

By de Broglie's relation, then, the momentum of the electron will also have a spread – an uncertainty Δp . This is as expected from the Uncertainty Relation. It can be shown mathematically (proof omitted here) that the wave packet description together with de Broglie relation and Born's probability interpretation reproduce the Heisenberg's Uncertainty Relation exactly.

In Chapter 13, the de Broglie relation will be seen to justify Bohr's postulate on quantisation of angular momentum of electron in an atom.

Figure 12.8 shows a schematic diagram of (a) a localized wave packet (b) an extended wave with fixed wavelength.

Example 12.3 What is the de Broglie wavelength associated with (a) an electron moving with a speed of 5.4×10^6 m/s, and (b) a ball of mass 150 g travelling at 30.0 m/s?

Answer

(a) For the electron:

Mass $m = 9.11 \times 10^{-31}$ kg, velocity $v = 5.4 \times 10^6$ m/s.

Then, momentum $p = mv = 9.11 \times 10^{-31}$ (kg) $\times 5.4 \times 10^6$ (m/s)

$$p = 4.92 \times 10^{-24} \text{ kg m/s}$$

de Broglie wavelength, $\lambda = h/p$

$$= \frac{6.63 \times 10^{-34} \text{ Js}}{4.92 \times 10^{-24} \text{ kg m/s}}$$

$$\lambda = 0.135 \text{ nm}$$

(b) For the ball:

Mass $m' = 0.150$ kg, velocity $v' = 30.0$ m/s.

Then momentum $p' = m' v' = 0.150$ (kg) $\times 30.0$ (m/s)

$$p' = 4.50 \text{ kg m/s}$$

de Broglie wavelength $\lambda' = h/p'$.

$$= \frac{6.63 \times 10^{-34} \text{ Js}}{4.50 \text{ kg m/s}}$$

$$\lambda' = 1.47 \times 10^{-34} \text{ m}$$

The de Broglie wavelength of electron is comparable with X-ray wavelengths. However, for the ball it is about 10^{-19} times the size of the proton, quite beyond experimental measurement.

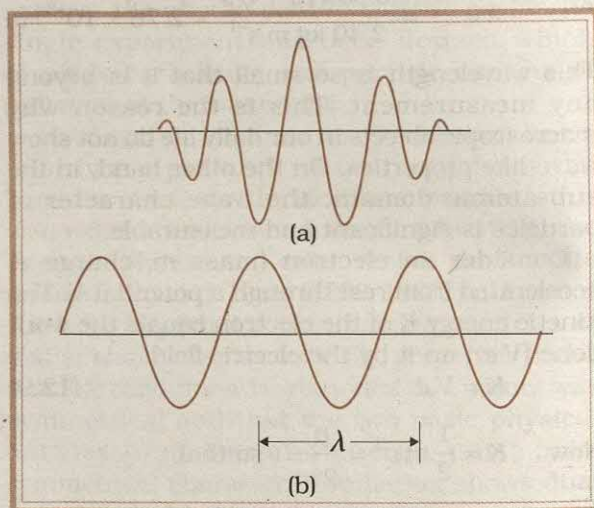


Fig. 12.8 (a) The wave packet description of an electron. The square of the amplitude of a point is related to probability density of the electron at that point. The wave packet corresponds to a spread of wavelength around some central wavelength (and hence by de Broglie relation, a spread in momentum). Consequently, it is associated with an uncertainty in position (Δx) and an uncertainty in momentum (Δp). (b) The matter wave corresponding to a definite momentum of an electron extends all over space. In this case, $\Delta p = 0$ and $\Delta x \rightarrow \infty$.

Example 12.4 What is the de Broglie wavelength associated with an electron, accelerated through a potential of 100 volts?

Answer Accelerating potential $V = 100$ V. The de Broglie wavelength λ is

$$\lambda = h/p = \frac{1.227}{\sqrt{V}} \text{ nm}$$

$$\lambda = \frac{1.227}{\sqrt{100}} \text{ nm} = 0.123 \text{ nm}$$

The de Broglie wavelength associated with an electron is of the order of X-ray wavelengths. ◀

12.10 THE DAVISSON AND GERMER EXPERIMENT

The wave nature of electrons was first experimentally verified by C.J. Davisson and L.H. Germer, in 1927, and independently by G.P. Thomson, in 1928, who observed diffraction effects with beams of electrons scattered by crystals. C.J. Davisson (1881-1958) and G.P. Thomson (1892-1975) shared the Nobel Prize in 1937 for their experimental discovery of diffraction of electrons by crystals.

The experimental arrangement used by Davisson and Germer is schematically shown in Fig. 12.9. It consists of an electron gun which comprises of a tungsten filament, coated with barium oxide and heated by a low tension (L.T.) battery. Electrons emitted by the filament are accelerated to a desired velocity by applying suitable potential from a high tension (H.T.) battery. They are collimated to a fine beam by allowing them to pass through a cylinder with fine holes along its axis. The fine collimated beam is made to fall on the surface of a nickel crystal. The electrons are scattered in all directions by the atoms of the crystal. The intensity of the electron beam, scattered in a given direction, is measured by the electron detector (collector). The detector can be rotated on a circular scale and is connected to a sensitive galvanometer, which records the current. The deflection of the galvanometer is proportional to the intensity of the electron beam entering the collector. The apparatus is enclosed in an evacuated chamber. By rotating the detector on the circular scale at different positions, the intensity of the scattered electron beam is measured for different

values of latitude angle (or angle of scattering) θ which is the angle between the incident and the scattered electron beams. The variation of the intensity (I) of the scattered electrons with the angle of scattering θ is obtained for different accelerating voltages.

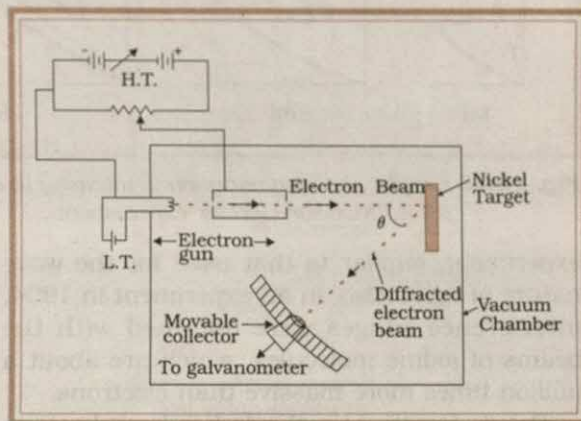


Fig. 12.9 Davisson-Germer electron diffraction arrangement.

Fig. 12.10 (a)–(e) show the results of Davisson and Germer experiment, when the accelerating voltage ranged from 44 V to 68 V. Note the appearance of a strong peak corresponding to a sharp diffraction maximum in the electron distribution at an accelerating potential of 54 V and scattering angle $\theta = 50^\circ$. The appearance of the peak in a particular direction is due to the constructive interference of electrons scattered from different layers of the regularly spaced atoms of the crystals. From the electron diffraction measurements, the wavelength of matter waves was found to be 0.165 nm.

The de Broglie wavelength λ associated with electron, using Eq. (12.12), for $V = 54$ V is given by

$$\lambda = h/p = \frac{1.227}{\sqrt{V}} \text{ nm}$$

$$\lambda = \frac{1.227}{\sqrt{54}} \text{ nm} = 0.166 \text{ nm}$$

There is thus an excellent agreement between the theoretical value and the experimentally obtained value of de Broglie wavelength. Davisson-Germer experiment thus strikingly confirms the wave nature of electrons and the de Broglie relation. More recently, in 1989, the wave nature of a beam of electrons was experimentally demonstrated in a double-slit

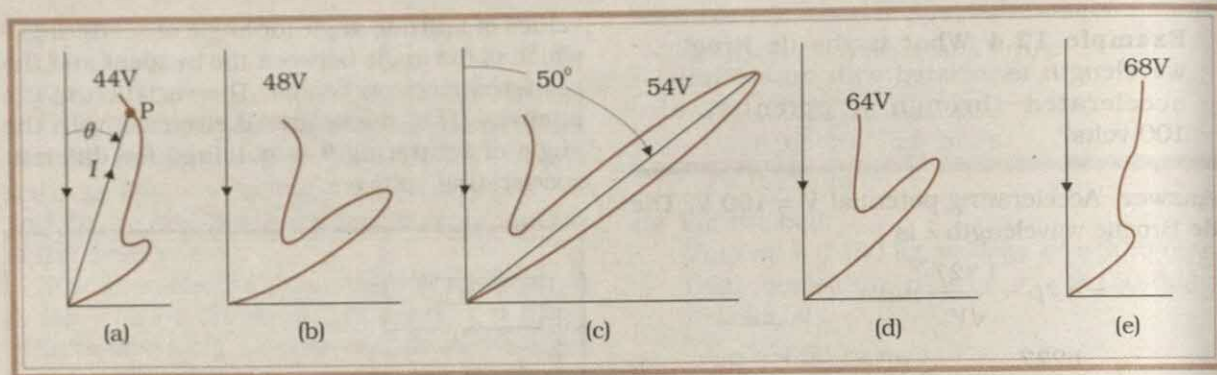


Fig. 12.10 Graphs showing variation in intensity (I) and scattering angle (θ) for different accelerating voltages (V) in Davisson-Germer experiment.

experiment, similar to that used for the wave nature of light. Also, in an experiment in 1994, interference fringes were obtained with the beams of iodine molecules, which are about a million times more massive than electrons.

The de Broglie hypothesis has been basic to the development of modern quantum mechanics. It has also led to the field of electron optics. The wave properties of electrons have been utilised in the design of electron microscope which is a great improvement, with higher resolution, over the optical microscope.

12.11 THE ELECTRON MICROSCOPE

The resolving limit of a microscope is the least distance between two point objects which can be distinguished (Chapter 11). The smallest resolvable separation of two points for an unaided eye is about 10^{-4} m or 100 μ m. A microscope using visible light gives the resolving limit to be $\lambda/2$ or about 250 nm. This is 400 times smaller than the resolving limit of the eye. Hence, the useful magnification of an optical microscope is limited to about 400. Greater magnification may make viewing by the microscope more comfortable but it will not reveal any additional significant details. As the diameter of a typical bacterial cell is about 1 μ m, it is not possible to make very detailed investigations of minute bacterial structures with an optical microscope. By using short wavelength ultraviolet light and a quartz optical system, we can improve the resolution and hence the useful magnification. However, for most optical work, the resolution limit is about 200 nm only.

The wave nature of electrons affords us the possibility of having probes of very short wavelength. Electrons sped up to high energies, using an accelerating voltage of, say, 50 kV have a de Broglie wavelength of 0.0055 nm, as can be verified using Eq. (12.12). This is about 10^5 times smaller than that of visible light. An electron microscope is a device that exploits the wave nature of electrons. Theoretically, the resolving limit of the electron microscope, using electrons of 50 keV, would be 0.0055 nm. However, in practice, the electron beam needs to be focussed using electric and magnetic fields as lenses (much like a beam of light is focussed using optical lenses). These limit the resolution to about 0.2 nm, which is still 1000 times better than that of the optical microscopes. The improved resolution makes it possible to investigate minute cellular constituents and molecules.

Fig. 12.11(a) gives the schematic diagram of an electron microscope that uses magnetic lenses to control the path of the electrons. The equivalent optical system is also shown in Fig. 12.11(b) for comparison. The electrons emitted from the hot filament source are accelerated by a high potential. This beam of electrons is further concentrated by a condenser magnetic lens (through the doughnut shaped electromagnet, similar to the condenser lens in an optical microscope). The magnetic lens is formed by passing suitable steady current through a coil of wire, enclosed in an iron shield, except for a gap of special design, which produces magnetic field of definite configuration. The object under investigation is placed on a thin collodion film in a special holder. The

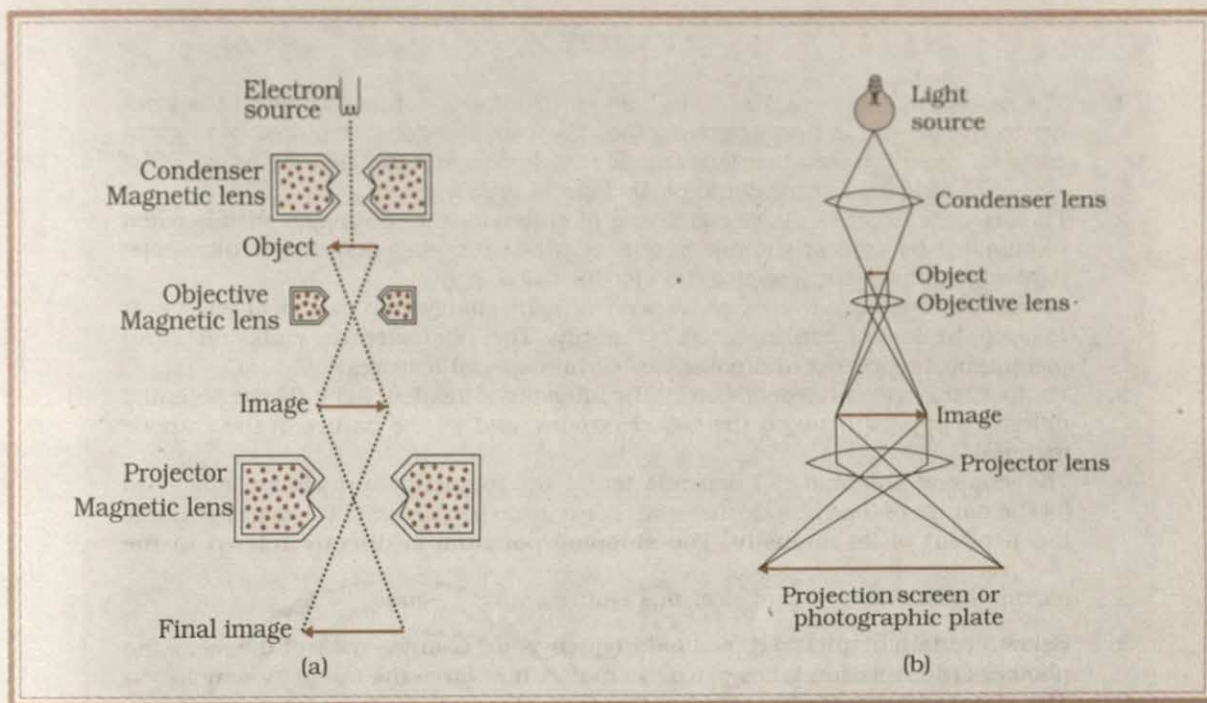


Fig. 12.11 (a) Electron microscope (b) Optical microscope.

concentrated beam of electrons strikes the object to be magnified and the electrons are partially absorbed by the object.

Below the object at a proper distance, an objective magnetic lens, similar to the objective in the optical microscope, is placed which spreads the electron beam and so produces the first enlarged image of the object. This image serves as a virtual object for the projector magnetic lens, similar to the eye-piece in an optical microscope. The projector magnetic lens produces the final enlarged image of the object on a fluorescent screen for visual examination or on a photographic plate for permanent record. The final image presents a shadow graph, similar to that produced by X-rays. The whole system is contained in an evacuated chamber (with very low pressure by means of a diffusion pump) so that the electrons are able to move freely from the filament to the screen.

For obtaining good photographs, free from aberrations as in an optical microscope, the accelerating potential and the currents used in the magnetic lenses must be kept steady. Accelerating voltages, of the order of a million volt, are usually employed. Using the electron

microscope, it has been possible to achieve magnification of about 10^6 (a million), with adequate resolving power, and particles of diameter less than 10 nm can be photographed.

Note that in the electron microscope, the electrons in a beam can be analysed as particles when their passage through the lens is concerned. Their wave nature comes into play only in establishing the resolution that the microscope can achieve.

The electron microscope, with its high magnifying and resolving powers, is one of the most indispensable and powerful tools for research in science, medicine and industry. It is a versatile device for investigating the atomic structure and crystal structure; the structural details of textile fibres, surfaces, colloids and polymers; composition of paper, paints and plastics; purification of lubricating oils, etc. It is useful in the selection and processing of ores, surface mapping and analysis of unknown materials. It has successfully been employed in the study of viruses (the disease causing agents) and the structural details of bacteria which, as mentioned before, cannot be resolved by ordinary optical microscopes.

SUMMARY

1. The minimum energy needed to pull an electron from a metal is called the work function of the metal. Energy (greater than the work function ϕ_0) required for electron emission from the metal surface can be supplied by suitably heating or, applying strong electric field or irradiating it by light of high frequency.
2. Photoelectric effect is the phenomenon of emission of electrons by metals when illuminated by light of suitable frequency. Certain metals respond to ultraviolet light while others are sensitive even to the visible light.

Photoelectric effect involves conversion of light energy into electrical energy. It follows the law of conservation of energy. The photoelectric emission is an instantaneous process and possesses certain special features.

3. Photoelectric current depends on (a) the intensity of incident light, (b) the potential difference applied between the two electrodes, and (c) the nature of the cathode material.
4. The stopping potential (V_0) depends on (a) the frequency of incident light, and (b) the nature of the cathode material. For a given frequency of incident light, it is independent of its intensity. The stopping potential is directly related to the

maximum kinetic energy of electrons emitted: $e V_0 = \frac{1}{2} m v_{\max}^2 = K_{\max}$.

5. Below a certain frequency (threshold frequency) ν_0 , characteristic of the metal, no photoelectric emission takes place, no matter how large the intensity may be.
6. The classical wave theory could not explain the main features of photoelectric effect. Its picture of continuous absorption of energy from radiation could not explain the independence of K_{\max} on intensity, the existence of ν_0 and the instantaneous nature of the process. Einstein explained these features on the basis of photon picture of light. According to this, light is composed of discrete packets of energy called quanta or photons. Each photon carries an energy $E (= h \nu)$ and momentum $p (= h/\lambda)$, which depend on the frequency (ν) of incident light and not on its intensity. Photoelectric emission from the metal surface occurs due to absorption of a photon by an electron.

7. Einstein's photoelectric equation is in accordance with the energy conservation law as applied to the photon absorption by an electron in the metal. The maximum

kinetic energy ($\frac{1}{2} m v_{\max}^2$) is equal to the photon energy ($h\nu$) minus the work function $\phi_0 (= h\nu_0)$ of the target metal:

$$\frac{1}{2} m v_{\max}^2 = V_0 e = h\nu - \phi_0 = h(\nu - \nu_0)$$

This photoelectric equation explains all the features of the photoelectric effect. Millikan's first precise measurements confirmed the Einstein's photoelectric equation and obtained an accurate value of Planck's constant h . This led to the acceptance of particle or photon description (nature) of electromagnetic radiation, introduced by Einstein.

8. A photo-cell employs photoelectric effect to convert change in intensity of illumination into a change in photoelectric current. Photo-cells are used to operate various control systems and in light measuring devices.
9. Radiation has dual nature: wave and particle. The nature of experiment determines whether a wave or particle description is best suited for understanding the experimental result. Reasoning that radiation and matter should be symmetrical in nature, Louis Victor de Broglie attributed a wave-like character to matter (material particles). The waves associated with the moving material particles are called matter waves or de Broglie waves.

10. The de Broglie wavelength (λ) associated with a moving particle is related to its momentum p as: $\lambda = h/p$. The dualism of matter is inherent in the de Broglie relation which contains a wave concept (λ) and a particle concept (p). The de Broglie wavelength is independent of the charge and nature of the material particle. It is significantly measurable (of the order of the atomic-planes spacing in crystals) only in case of sub-atomic particles like electrons, protons, etc. (due to smallness of their masses and hence, momenta). However, it is indeed very small, quite beyond measurement, in case of macroscopic objects, commonly encountered in everyday life.
11. Electron diffraction experiments by Davisson and Germer, and by G. P. Thomson, as well as many later experiments, have verified and confirmed the wave-nature of electrons. The de Broglie hypothesis of matter waves supports the Bohr's concept of stationary orbits.
12. Electron microscope is a device that exploits the wave-nature of electrons to provide high resolving power. It has successfully been employed to investigate structural details of bacteria, viruses, etc. It has proved to be a powerful tool of investigation for research in science, technology, metallurgy, industry, medicine, etc.

Physical Quantity	Symbol	Dimensions	Unit	Remarks
Planck's constant	h	$[ML^2T^{-1}]$	J s	$E = h\nu$
Stopping potential	V_0	$[ML^2T^{-3}A^{-1}]$	V	$eV_0 = K_{\max}$
Work function	ϕ_0	$[ML^2T^{-2}]$	J; eV	$K_{\max} = E - \phi_0$
Threshold frequency	ν_0	$[T^{-1}]$	Hz	$\nu_0 = \phi_0/h$
de Broglie wavelength	λ	[L]	m	$\lambda = h/p$

POINTS TO PONDER

1. Free electrons in a metal are free in the sense that they move inside the metal in a constant potential. (This is only an approximation.) They are not free to move out of the metal. They need additional energy to get out of the metal.
2. Free electrons in a metal do not all have the same energy. Like molecules in a gas jar, the electrons have a certain energy distribution at a given temperature. This distribution is different from the usual Maxwell's distribution that you have learnt in the study of kinetic theory of gases. You will learn about it in later courses, but the difference has to do with the fact that electrons obey Pauli's exclusion principle.
3. Because of the energy distribution of free electrons in a metal, the energy required by an electron to come out of the metal is different for different electrons. Electrons with higher energy require less additional energy to come out of the metal than those with lower energies. Work function is the least energy required by an electron to be out of the metal.
4. Observations on photoelectric effect imply only that in interaction of light with electron, **absorption of energy takes place in discrete units of $h\nu$** . This is not quite the same as saying that light consists of particles, each of energy $h\nu$.

5. Observations on the stopping potential (its independence of intensity and dependence on frequency) are the crucial discriminator between the wave-picture and photon-picture of photoelectric effect.
6. The wavelength of a matter wave given by $\lambda = \frac{h}{p}$ has physical significance; its phase velocity v_p has no physical significance. However, the group velocity of the matter wave is physically meaningful and equals the velocity of the particle.

EXERCISES

- 12.1** Find the
 - (a) maximum frequency, and
 - (b) minimum wavelength of X-rays produced by 30 kV electrons.
- 12.2** The work function of caesium metal is 2.14 eV. When light of frequency 6×10^{14} Hz is incident on the metal surface, photoemission of electrons occurs. What is the
 - (a) maximum kinetic energy of the emitted electrons,
 - (b) Stopping potential, and
 - (c) maximum speed of the emitted photoelectrons?
- 12.3** The photoelectric cut-off voltage in a certain experiment is 1.5 V. What is the maximum kinetic energy of photoelectrons emitted?
- 12.4** Monochromatic light of wavelength 632.8 nm is produced by a helium-neon laser. The power emitted is 9.42 mW.
 - (a) Find the energy and momentum of each photon in the light beam,
 - (b) How many photons per second, on the average, arrive at a target irradiated by this beam? (Assume the beam to have uniform cross-section which is less than the target area), and
 - (c) How fast does a hydrogen atom have to travel in order to have the same momentum as that of the photon?
- 12.5** The energy flux of sunlight reaching the surface of the earth is 1.388×10^3 W/m². How many photons (nearly) per square metre are incident on the Earth per second? Assume that the photons in the sunlight have an average wavelength of 550 nm.
- 12.6** In an experiment on photoelectric effect, the slope of the cut-off voltage versus frequency of incident light is found to be 4.12×10^{-15} V s. Calculate the value of Planck's constant.
- 12.7** A 100 W sodium lamp radiates energy uniformly in all directions. The lamp is located at the centre of a large sphere that absorbs all the sodium light which is incident on it. The wavelength of the sodium light is 589 nm. (a) What is the energy per photon associated with the sodium light? (b) At what rate are the photons delivered to the sphere?
- 12.8** What are the energies of photons at the (a) violet, and (b) red ends of the visible spectrum? The wavelength of light is about 390 nm for violet and about 760 nm for red.
- 12.9** From which of the photosensitive materials with work function listed in Table 12.1 can you build a photo-cell that operates with visible light? (Use the results of Exercise 12.8.)

- 12.10** The threshold frequency for a certain metal is 3.3×10^{14} Hz. If light of frequency 8.2×10^{14} Hz is incident on the metal, predict the cut-off voltage for the photoelectric emission.
- 12.11** The work function for a certain metal is 4.2 eV. Will this metal give photoelectric emission for incident radiation of wavelength 330 nm?
- 12.12** Light of frequency 7.21×10^{14} Hz is incident on a metal surface. Electrons with a maximum speed of 6.0×10^5 m/s are ejected from the surface. What is the threshold frequency for photoemission of electrons?
- 12.13** Light of wavelength 488 nm is produced by an argon laser which is used in the photoelectric effect. When light from this spectral line is incident on the cathode, the stopping (cut-off) potential of photoelectrons is 0.38 V. Find the work function of the material from which the cathode is made.
- 12.14** Calculate the
 (a) momentum, and
 (b) de Broglie wavelength of the electrons accelerated through a potential difference of 56 V.
- 12.15** What is the
 (a) momentum,
 (b) speed, and
 (c) de Broglie wavelength of an electron with kinetic energy of 120 eV.
- 12.16** The wavelength of light from the spectral emission line of sodium is 589 nm. Find the kinetic energy at which
 (a) an electron, and
 (b) a neutron, would have the same de Broglie wavelength.
- 12.17** What is the de Broglie wavelength of
 (a) a bullet of mass 0.040 kg travelling at the speed of 1.0 km/s,
 (b) a ball of mass 0.060 kg moving at a speed of 1.0 m/s, and
 (c) a dust particle of mass 1.0×10^{-9} kg drifting with a speed of 2.2 m/s?
- 12.18** An electron and a photon each have a wavelength of 1.00 nm. Find
 (a) their momenta,
 (b) the energy of the photon, and
 (c) the kinetic energy of electron.
- 12.19** (a) For what kinetic energy of a neutron will the associated de Broglie wavelength be 1.40×10^{-10} m?
 (b) Also find the de Broglie wavelength of a neutron, in thermal equilibrium with matter, having an average kinetic energy of $\frac{3}{2} kT$ at 300 K.*
- 12.20** Show that the wavelength of electromagnetic radiation is equal to the de Broglie wavelength of its quantum (photon).
- 12.21** An electron, α -particle, and a proton have the same kinetic energy. Which of these particles has the shortest de Broglie wavelength?
- 12.22** A particle is moving three times as fast as an electron. The ratio of the de Broglie wavelength of the particle to that of the electron is 1.813×10^{-4} . Calculate the particle's mass and identify the particle.
- 12.23** What is the de Broglie wavelength of a nitrogen molecule in air at 300 K? Assume that the molecule is moving with the root-mean-square speed of molecules at this temperature. (Atomic mass of nitrogen = 14.0076 u)

ADDITIONAL EXERCISES

- 12.24** (a) Estimate the speed with which electrons emitted from a heated cathode of an evacuated tube impinge on the anode maintained at a potential difference of 500 V with respect to the cathode. Ignore the small initial speeds of the electrons. The 'specific charge' of the electron i.e., its e/m is given to be $1.76 \times 10^{11} \text{ C kg}^{-1}$.
- (b) Use the same formula you employ in (a) to obtain electron speed for an anode potential of 10 MV. Do you see what is wrong? In what way is the formula to be modified?
- 12.25** (a) A monoenergetic electron beam with electron speed of $5.20 \times 10^6 \text{ m s}^{-1}$ is subject to a magnetic field of $1.30 \times 10^{-4} \text{ T}$ normal to the beam velocity. What is the radius of the circle traced by the beam, given e/m for electron equals $1.76 \times 10^{11} \text{ C kg}^{-1}$.
- (b) Is the formula you employ in (a) valid for calculating radius of the path of a 20 MeV electron beam? If not, in what way is it modified?
- [Note: Exercises 12.24(b) and 12.25(b) take you to relativistic mechanics which is beyond the scope of this book. They have been inserted here simply to emphasise the point that the formulas you use in part (a) of the exercises are not valid at very high speeds or energies. See answers at the end to know what 'very high speed or energy' means.]
- 12.26** In a Thomson's set-up for determining e/m , the same high tension dc supply provides potential to the anode of the accelerating column, as also to the positive deflecting plate in the region of crossed fields. If the supply voltage is doubled, by what factor should the magnetic field be increased to keep the electron beam undeflected?
- 12.27** The deflecting plates in a Thomson's set-up are 5.0 cm long, and 1.5 cm apart. The plates are maintained at a potential difference of 240 V. Electrons accelerated to an energy of 2.0 keV enter from one edge of the plates midway in a direction parallel to the plates.
- (a) What is the deflection at the other edge of the plates?
- (b) At what distance from the undeflected position of the screen does the beam strike if the screen is 30 cm away from the other edge of the plates?
- 12.28** In a Thomson's set-up for determination of e/m , a uniform electric field $E = 24.0 \text{ kV m}^{-1}$ set up between two parallel plates of length 6.0 cm each produces a deflection of 10.9 cm on the fluorescent screen. A magnetic field is then switched on and adjusted to the value $B = 8.0 \times 10^{-4} \text{ T}$ to restore the beam to its undeflected position. The distance of the screen from the centre of the plates is 40.0 cm. Determine e/m from the data.
- 12.29** An electron gun with its anode at a potential of 100 V fires out electrons in a spherical bulb containing hydrogen gas at low pressure ($\sim 10^{-2} \text{ mm of Hg}$). A magnetic field of $2.83 \times 10^{-4} \text{ T}$ curves the path of the electrons in a circular orbit of radius 12.0 cm. (The path can be viewed because the gas ions in the path focus the beam by attracting electrons, and emitting light by electron capture; this method is known as the 'fine beam tube' method.) Determine e/m from the data.
- 12.30** In a Millikan's oil-drop experiment, a charged oil drop of mass density 880 kg m^{-3} is held stationary between two parallel plates 6.00 mm apart held at a potential difference of 10^3 V . When the electric field is switched off, the drop is observed to fall a distance of 2.00 mm in 35.7 s.
- (a) What is the radius of the drop?
- (b) Estimate the charge of the drop. How many excess electrons does it carry?

(The upper plate in the experiment is at a higher potential.) (Viscosity of air = $1.80 \times 10^{-5} \text{ N s m}^{-2}$; $g = 9.81 \text{ m s}^{-2}$; density of air = 1.29 kg m^{-3})

12.31 (Millikan's experiment not only measured charge of an electron; it also established a fundamental fact of nature: charge is quantised i.e., charge comes in multiples of a basic unit, namely, the electronic charge e . Make sure you understand this point by going through the following exercise): In a Millikan's oil-drop set-up, an oil-drop falls with a terminal speed v_0 in the absence of any electric field. When a fixed electric field is switched on, and a radioactive source is kept in the environment, it is found that the same oil drop, when viewed for long, shows up different terminal speeds v_1, v_2, v_3, \dots

- What causes the drop to change its terminal speed in the same electric field? (Assume size and mass of the drop remain unchanged).
- What key observation on the different terminal speeds of the drop suggests charge quantisation?

12.32 In a variant of the Millikan's oil-drop set-up, an oil drop whose radius is measured by a separate observation to be $1.0 \times 10^{-6} \text{ m}$, falls down in the absence of any electric field with a certain terminal velocity. When a horizontal electric field is set up by means of two parallel vertical plates held 10 mm apart at a potential difference of 1500 V, the drop is seen to fall steadily at an angle of 63° with the vertical. The density of the oil used is 900 kg m^{-3} . Estimate the charge on the drop.

- 12.33** (a) An X-ray tube produces a continuous spectrum of radiation with its short wavelength end at 0.45 \AA . What is the maximum energy of a photon in the radiation?
- (b) From your answer to (a), guess what order of accelerating voltage (for electrons) is required in such a tube?

12.34 In an accelerator experiment on high-energy collisions of electrons with positrons, a certain event is interpreted as annihilation of an electron-positron pair of total energy 10.2 BeV into two γ -rays of equal energy. What is the wavelength associated with each γ -ray? ($1 \text{ BeV} = 10^9 \text{ eV}$)

12.35 Estimating the following two numbers should be interesting. The first number will tell you why radio engineers do not need to worry much about photons! The second number tells you why our eye can never 'count photons', even in barely detectable light.

- The number of photons emitted per second by a Medium wave transmitter of 10 kW power, emitting radiowaves of wavelength 500 m.
- The number of photons entering the pupil of our eye per second corresponding to the minimum intensity of white light that we humans can perceive ($\sim 10^{-10} \text{ W m}^{-2}$). Take the area of the pupil to be about 0.4 cm^2 , and the average frequency of white light to be about $6 \times 10^{14} \text{ Hz}$.

12.36 Ultraviolet light of wavelength 2271 \AA from a 100 W mercury source irradiates a photo-cell made of molybdenum metal. If the stopping potential is -1.3 V , estimate the work function of the metal. How would the photo-cell respond to a high intensity ($\sim 10^5 \text{ W m}^{-2}$) red light of wavelength 6328 \AA produced by a He-Ne laser?

12.37 Monochromatic radiation of wavelength 640.2 nm ($1 \text{ nm} = 10^{-9} \text{ m}$) from a neon lamp irradiates photosensitive material made of caesium on tungsten. The stopping voltage is measured to be 0.54 V . The source is replaced by an iron source and its 427.2 nm line irradiates the same photo-cell. Predict the new stopping voltage.

- 12.38** A mercury lamp is a convenient source for studying frequency dependence of photoelectric emission, since it gives a number of spectral lines ranging from the UV to the red end of the visible spectrum. In our experiment with rubidium photo-cell, the following lines from a mercury source were used:

$$\lambda_1 = 3650 \text{ \AA}, \lambda_2 = 4047 \text{ \AA}, \lambda_3 = 4358 \text{ \AA}, \lambda_4 = 5461 \text{ \AA}, \lambda_5 = 6907 \text{ \AA},$$

The stopping voltages, respectively, were measured to be:

$$V_{01} = 1.28 \text{ V}, V_{02} = 0.95 \text{ V}, V_{03} = 0.74 \text{ V}, V_{04} = 0.16 \text{ V}, V_{05} = 0 \text{ V}$$

(a) Determine the value of Planck's constant h .

(b) Estimate the threshold frequency and work function for the material.

[Note: You will notice that to get h from the data, you will need to know e (which you can take to be $1.6 \times 10^{-19} \text{ C}$). Experiments of this kind on Na, Li, K, etc. were performed by Millikan, who, using his own value of e (from the oil-drop experiment) confirmed Einstein's photoelectric equation and at the same time gave an independent estimate of the value of h .]

- 12.39** The work function for the following metals is given:

Na: 2.75 eV; K: 2.30 eV; Mo: 4.17 eV; Ni: 5.15 eV. Which of these metals will not give photoelectric emission for a radiation of wavelength 3300 Å from a He-Cd laser placed 1 m away from the photocell? What happens if the laser is brought nearer and placed 50 cm away?

- 12.40** Light of intensity 10^{-5} W m^{-2} falls on a sodium photo-cell of surface area 2 cm^2 . Assuming that the top 5 layers of sodium absorb the incident energy, estimate time required for photoelectric emission in the wave-picture of radiation. The work function for the metal is given to be about 2 eV. What is the implication of your answer?

- 12.41** (a) Show that a free electron at rest cannot absorb a photon and thereby acquire kinetic energy equal to the energy of the photon. Would the conclusion change if the free electron was moving with a constant velocity?

(b) If the absorption of a photon by a free electron is ruled out as proved in (a) above, how does photoelectric emission take place at all?

- 12.42** In an experiment on photoelectric emission by γ -rays on platinum, the energy distribution of photoelectrons exhibits peaks at a number of discrete energies: 270 keV, 339 keV and 354 keV. The binding energies of K, L and M shells in platinum are known to be 77 keV, 13 keV and 3.5 keV approximately. What is the wavelength of the γ -rays with which the data are consistent?

- 12.43** An X-ray pulse is sent through a section of Wilson cloud chamber containing a supersaturated gas, and tracks of photoelectrons ejected from the gaseous atoms are observed. Two groups of tracks of lengths 1.40 cm and 2.02 cm are noted. If the range-energy relation for the cloud chamber is given by $R = \alpha E$ with $\alpha = 1 \text{ cm/keV}$, obtain the binding energies of the two levels from which electrons are emitted. (Wavelength of the X-rays pulse = 4.9 Å)

[Note: Exercises 12.42 and 12.43 take you away from the typical scenario of photoelectric effect that you have learnt in this chapter. Actually, like visible and UV radiation, X-rays and γ -rays also cause photoelectric emission. But their photons have much greater energies, and so can eject electrons of much greater binding energies than those ejected by visible or UV. They can eject electrons from the inner shells of individual atoms where the energies are discrete. The energy distribution of electrons emitted by X-ray or γ -rays photoelectric effect may, consequently, exhibit sharp peaks at discrete energies (at the lower end of the energy spectrum).]

- 12.44** Crystal diffraction experiments can be performed using X-rays, or electrons accelerated through appropriate voltage. Which probe has greater energy? (For quantitative comparison, take the wavelength of the probe equal to 1 Å, which is of the order of inter-atomic spacing in the lattice) ($m_e = 9.11 \times 10^{-31}$ kg).
- 12.45** (a) Obtain the de Broglie wavelength of a neutron of kinetic energy 150 eV. As you have seen in Exercise 12.44, an electron beam of this energy is suitable for crystal diffraction experiments. Would a neutron beam of the same energy be equally suitable? Explain. ($m_n = 1.675 \times 10^{-27}$ kg).
 (b) Obtain the de Broglie wavelength associated with thermal neutrons at room temperature (27 °C). Hence explain why a fast neutron beam needs to be thermalised with the environment before it can be used for neutron diffraction experiments.
- 12.46** An electron microscope uses electrons accelerated by a voltage of 50 kV. Determine the de Broglie wavelength associated with the electrons. If other factors (such as numerical aperture, etc.) are taken to be roughly the same, how does the resolving power of an electron microscope compare with that of an optical microscope which uses yellow light?
- 12.47** The wavelength of a probe is roughly a measure of the size of a structure that it can probe in some detail. The quark structure of protons and neutrons appears at the minute length-scale of 10^{-15} m or less. This structure was first probed in early 1970's using high energy electron beams produced by a linear accelerator at Stanford, USA. Guess what might have been the order of energy of these electron beams. (Rest mass energy of electron = 0.511 MeV.)
- 12.48** The extent of localisation of a particle is determined roughly by its de Broglie wavelength. If an electron is localised within the nucleus (of size about 10^{-14} m) of an atom, what is its energy? Compare this energy with the typical binding energies (of the order of a few MeV) in a nucleus, and hence argue why electrons cannot reside in a nucleus.
- 12.49** Find the typical de Broglie wavelength associated with a He atom in helium gas at room temperature (27 °C) and 1 atm pressure; and compare it with the mean separation between two atoms under these conditions.
- 12.50** Compute the typical de Broglie wavelength of an electron in a metal at 27 °C and compare it with the mean separation between two electrons in a metal which is given to be about 2×10^{-10} m.
 [Note: Exercises 12.49 and 12.50 reveal that while the wave-packets associated with gaseous molecules under ordinary conditions are non-overlapping, the electron wave-packets in a metal strongly overlap with one another. This suggests that whereas molecules in an ordinary gas can be distinguished apart, electrons in a metal cannot be distinguished apart from one another. This indistinguishability has many fundamental implications which you will explore in more advanced Physics courses.]
- 12.51** Answer the following questions:
 (a) Quarks inside protons and neutrons are thought to carry fractional charges [$(+2/3)e$; $(-1/3)e$]. Why do they not show up in Millikan's oil-drop experiment?
 (b) Why do we need the oil-drops of Millikan's experiment to be of such microscopic sizes? Why cannot we experiment with much bigger drops?
 (c) Stokes' formula for viscous drag is not really valid for oil-drops of extremely minute sizes. Why not?
 (d) What is so special about the combination e/m ? Why do we not simply talk of e and m separately?

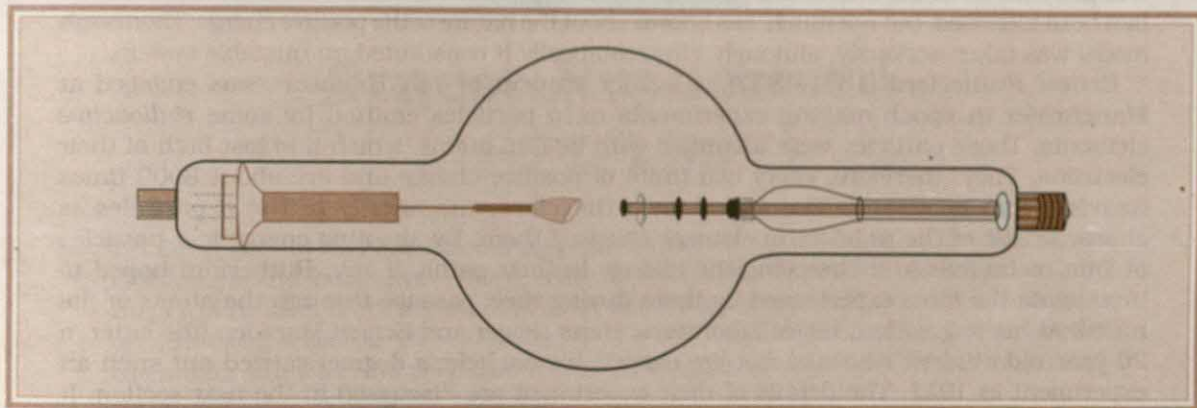
- (e) Why should gases be insulators at ordinary pressures and start conducting at very low pressures?
- (f) Every metal has a definite work function. Why do all photoelectrons not come out with the same energy if incident radiation is monochromatic? Why is there an energy distribution of photoelectrons?
- (g) The energy and momentum of an electron are related to the frequency and wavelength of the associated matter wave by the relations:

$$E = h\nu, p = \frac{h}{\lambda}$$

But while the value of λ is physically significant, the value of ν (and therefore, the value of the phase speed $\nu\lambda$) has no physical significance. Why?

CHAPTER THIRTEEN

ATOMS



13.1 INTRODUCTION

In the nineteenth century, enough evidence had accumulated in favour of the hypothesis that each element has its distinctive atom. In 1897, the experiments on electric discharge through gases carried out by the British physicist J.J. Thomson revealed that atoms of different elements contain negatively charged constituents (electrons) that are completely identical. At this stage, the specific charge (e/m) of an electron was known but its mass was not known. Thus, it was not possible even to say how many negatively charged electrons a given atom contained. The electrical neutrality of atoms is, of course, evident, since even a small net charge on an atom would imply a large charge for a macroscopic body (containing atoms of the order of Avogadro's number), which is not observed. Therefore, an atom must also contain some positive charge. In the first years of twentieth century, nobody knew what form this compensating positive charge took.

The pioneering oil-drop experiment by the American physicist R.A. Millikan established the quantisation of electric charge. He demonstrated that the *elementary charge*, carried by an electron, is 1.6×10^{-19} C.

As of now 117 elements are known, of these 109 have been named. How do these elements differ from each other? It is reasonable to assume that the atoms of different elements differ from each other because they contain different numbers of electrons and positive charges. How are the electrons and positive charges distributed inside an atom? In other words, what is the structure of the atom? In this Chapter, we shall dwell upon this question.

In the mid-nineteenth century, it was speculated that the motion of electric charge within individual atoms produced visible light. In 1862, Faraday placed a light source in a strong magnetic field in an attempt to determine whether the field changed the emitted radiation.

He was not able to detect any change, but when Zeeman repeated his experiment thirty years later with improved equipment, changes were observed.

The thermal radiation from a hot body has a continuous spread of wavelengths. In contrast, light emitted from atoms heated in a flame, or excited electrically in a glow tube such as the familiar neon sign or mercury vapour light has only certain discrete wavelengths. In a spectrometer using a narrow slit in conjunction with a prism, the spectrum appears as a series of bright lines. It was learnt early in the nineteenth century that each element is associated with a characteristic spectrum of radiation. This fact suggested an intimate relationship between the internal structure of an atom and the spectrum of radiation emitted by it. In later sections of this Chapter, we shall see how the information on atomic spectra provides useful clues to the structure of an atom.

In 1898, J.J. Thomson proposed that an atom is basically a spherical cloud of positive charges with electrons embedded in it like seeds in a watermelon. At that time, the electrons had been identified, but not much was known about the nature of the positive charge. Thomson's model was taken seriously, although, electrostatically it constituted an unstable system.

Ernest Rutherford (1871-1937), a former student of J.J. Thomson, was engaged at Manchester in epoch-making experiments on α -particles emitted by some *radioactive* elements. These particles were identified with helium atoms, which had lost both of their electrons. They, therefore, carry two units of positive charge and are about 8000 times heavier than electrons. The speed and, therefore, the energy of the α -particles is characteristic of the radioactive element emitting them. By shooting energetic α -particles at thin metal foils and observing the change in their paths, if any, Rutherford hoped to investigate the force experienced by them during their passage through the atoms of the metal. At his suggestion, his collaborators, Hans Geiger and Ernest Marsden (the latter, a 20 year-old student who had not yet earned his bachelor's degree) carried out such an experiment in 1911. The details of their experiment are discussed in the next section. It was observed that although most of the α -particles travelled through the thin foil undeviated, some suffered a large change in direction and a few even deflected backwards.

Rutherford saw that, to deflect the α -particle backwards, it must experience a large force; this force could be provided if the positive charge, instead of being distributed uniformly throughout the atom, were concentrated tightly at its center. Then the incoming α -particle could get very close to the positive charge without penetrating it, and such a close encounter would result in a large deflection. This explanation led to the birth of Rutherford's planetary model of atom. In this model, the entire positive charge and most of the mass of the atom are concentrated in a small volume called the nucleus and electrons revolve around it just as the planets do around the sun. We next discuss the α -scattering experiment that led to this model.



Sir John Joseph Thomson

(1856-1940)

British scientist who investigated the nature of cathode rays. He discovered the electron by his ingenious experiments on discharge of electricity through gases.



Ernest Rutherford **(1871-1937)**

British physicist who did pioneering work on radioactive radiation. He discovered alpha-rays and beta-rays. Along with Frederick Soddy, he created the modern theory of radioactivity. He studied thorium emanation which led to the discovery of the noble gas thoron. By scattering alpha-rays from thin metal foils, he discovered the atomic nucleus and proposed the planetary model of atom. He also estimated the approximate size of the nucleus.



Niels Henrik David Bohr **(1885-1962)**

Danish physicist who explained the spectrum of hydrogen atom based on quantum ideas. He gave a theory of nuclear fission based on the liquid-drop model of nucleus. Bohr contributed to the clarification of conceptual problems in quantum mechanics, in particular by proposing the complementarity principle.

13.2 ALPHA-PARTICLE SCATTERING AND RUTHERFORD'S MODEL OF ATOM

At the suggestion of Rutherford, in 1911, H. Geiger and E. Marsden performed some classic experiments. In one of their experiments, 5.5 MeV α -particles emitted from $^{214}_{83}\text{Bi}$ radioactive source were scattered from a thin gold foil.

The schematic arrangement in the Geiger-Marsden experiment is shown in Fig. 13.1.

Alpha-particles emitted by a $^{214}_{83}\text{Bi}$ radioactive source were collimated into a narrow beam by their passage through lead bricks. The beam was allowed to fall on a thin foil of gold of thickness $2.1 \times 10^{-7}\text{m}$. The scattered alpha-particles were observed through a rotatable detector consisting of zinc sulphide screen and a microscope. The scattered alpha-particles on striking the screen produced bright light flashes or scintillations. These scintillations could be viewed through the microscope and counted at different angles from the direction of the incident beam.

A typical graph of the total number of alpha-particles scattered at different angles, in a given interval of time, is shown in Fig. 13.2. The dots in this figure represent the data points and the solid curve is the theoretical prediction based on the assumption that the atom has a small, massive, positively charged nucleus.

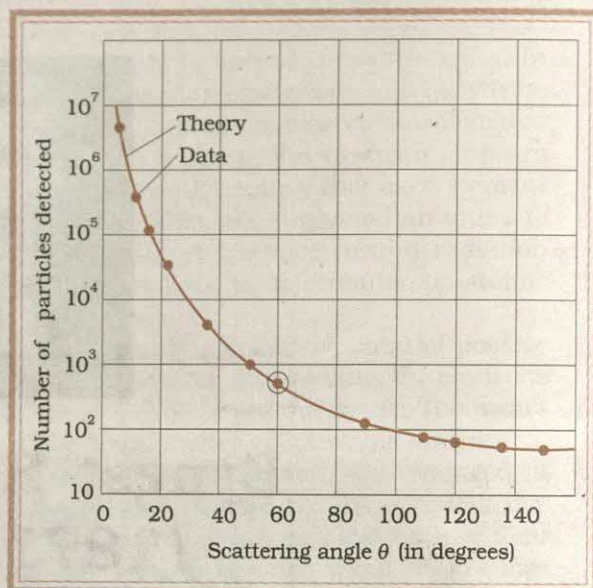


Fig. 13.2 The dots are alpha-particle scattering data for a gold foil, obtained by Geiger and Marsden using the set-up shown in Fig. 13.1. The solid curve is the theoretical prediction; based on the assumption that atom has a small, massive, positively charged nucleus. The data have been adjusted to fit the theoretical curve at a point enclosed by a circle.

The scattering data shown in Fig. 13.2 can be analysed by employing Rutherford's nuclear

model of the atom. As the gold foil is very thin, it can be assumed that alpha-particles will suffer not more than one scattering during their passage through it. Therefore, computation of the trajectory of an alpha-particle scattered by a single nucleus is enough. Alpha-particles are nuclei of helium atoms and, therefore, carry two units, $2e$, of positive charge and have the mass of the helium atom. The charge of the gold

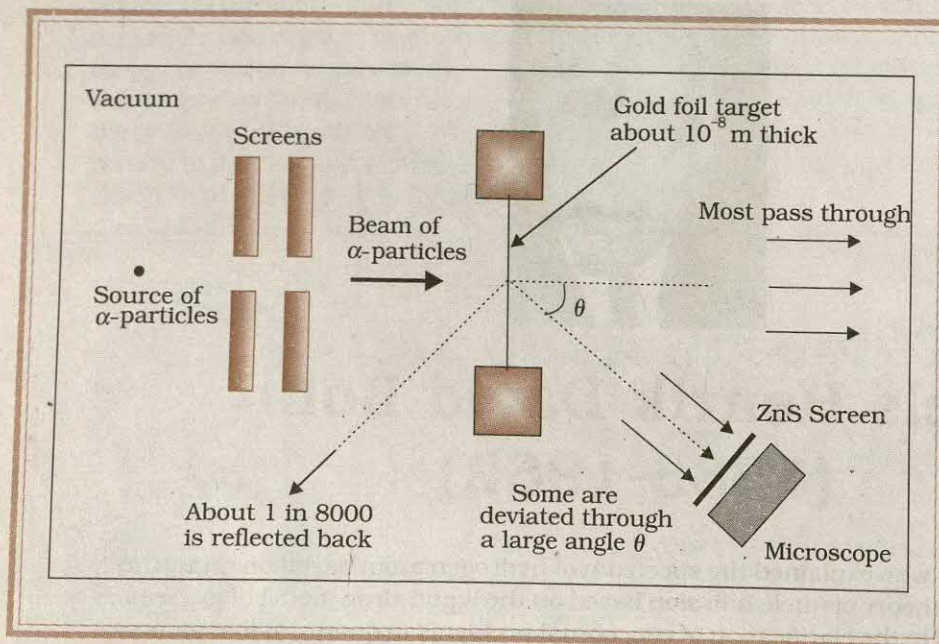


Fig. 13.1 Schematic arrangement of the Geiger-Marsden experiment.

nucleus is Ze , where Z^* is the atomic number of the atom; for gold $Z = 79$. Since the nucleus of gold is about 50 times heavier than an alpha-particle, it is reasonable to assume that it remains stationary throughout the scattering process. Under these assumptions, the trajectory of an alpha-particle can be computed employing Newton's second law of motion and the Coulomb's law for electrostatic force of repulsion between the alpha-particle and the positively charged nucleus. The magnitude of this force is given by,

$$F = \frac{1}{4\pi\epsilon_0} \frac{(2e)(Ze)}{r^2} \quad (13.1)$$

where r is the distance between the alpha-particle and the nucleus. The force is directed along the line joining the alpha-particle and the nucleus. The magnitude and direction of the force on an alpha-particle continuously changes as it approaches the nucleus and recedes away from it.

The trajectory traced by an alpha-particle depends on its distance of closest approach from a nucleus or an equivalent length called the *impact parameter*, b (Fig.13.3). It is the distance of the initial velocity vector of the alpha-particle from the centre of the nucleus. In Fig. 13.4, we have shown the numerically computed trajectories for four arbitrary values of the impact parameter; curve 1 is for the largest value of b and curve 4 for the smallest. For large impact parameters, the force experienced by the alpha-particle is weak because of its inverse square law character. It is seen that for a large impact parameter the alpha-particle goes nearly undeviated and for small impact parameter, it suffers large scattering. In case of head-on collision, the impact parameter is zero and the alpha-particle rebounds back. Rutherford had analytically calculated the relation between the impact parameter b and the scattering angle θ . It is

$$b = \frac{Ze^2 \cot(\theta/2)}{4\pi\epsilon_0 E} \quad (13.2)$$

where E is the kinetic energy of the incident alpha-particle. The derivation of this relation is outside the scope of the present text.

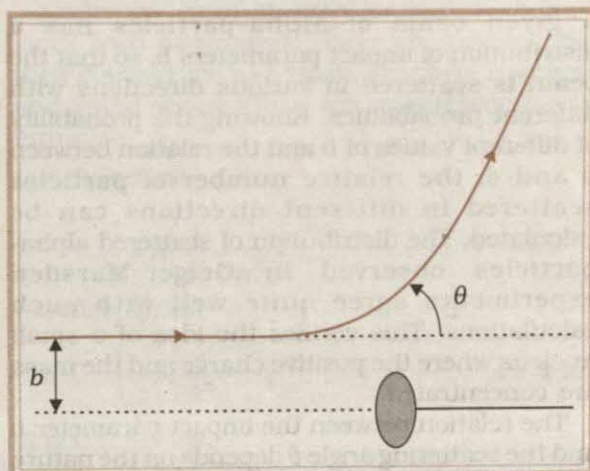


Fig. 13.3 Trajectory of an alpha-particle in the Coulomb field of a heavy nucleus. The impact parameter b and the scattering angle are defined as shown in the diagram.

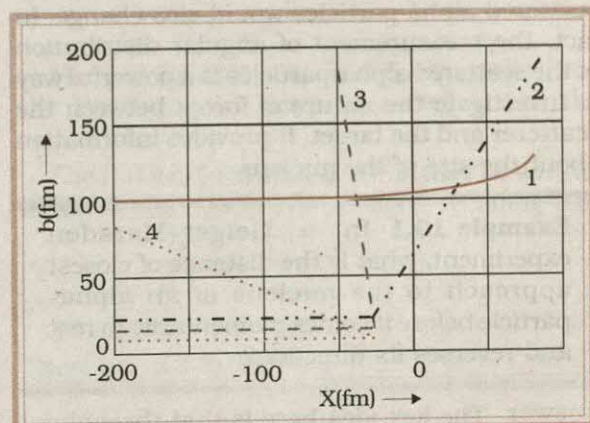


Fig. 13.4 Numerically computed trajectories of alpha-particles, moving with the speed of 1.63×10^7 m/s far away from the nucleus, in the Coulomb field of a gold nucleus. The gold nucleus is located at the origin of the coordinate system. The values of the impact parameters are 2.5 fm (curve 4), 10 fm (curve 3), 20 fm (curve 2), and 100 fm (curve 1).

The Eq. (13.2) shows that an alpha-particle close to the nucleus (small impact parameter) will have a large deflection ($\theta \approx \pi$) whereas an alpha-particle far away from the nucleus (large impact parameter) will have a small deflection ($\theta \approx 0$). We also see that if the kinetic energy of the alpha-particle is large, b can be small.

* The determination of Z (atomic number) will be discussed later.

A given beam of alpha-particles has a distribution of impact parameters b , so that the beam is scattered in various directions with different probabilities. Knowing the probability of different values of b and the relation between b and θ , the relative number of particles scattered in different directions can be calculated. The distribution of scattered alpha-particles observed in Geiger-Marsden experiments agree quite well with such calculations. This verifies the idea of a small nucleus where the positive charge and the mass are concentrated.

The relation between the impact parameter b and the scattering angle θ depends on the nature of the force law. For example, Eq. (13.2) holds if the nucleus is a point charge. But if the force law changes as the alpha-particles come closer to the nucleus (as is expected if the nucleus has a finite size), the angular distribution of the scattered alpha-particles would also change. In fact, the measurement of angular distribution of the scattered alpha-particles is a powerful way to investigate the nature of forces between the scatterer and the target. It provides information about the size of the nucleus.

Example 13.1 In a Geiger-Marsden experiment, what is the distance of closest approach to the nucleus of an alpha-particle before it comes momentarily to rest and reverses its direction?

Answer The key idea here is that throughout the scattering process, the total mechanical energy of the system consisting of an alpha-particle and a gold nucleus is conserved. The system's initial mechanical energy E_i before the particle and nucleus interact, is equal to its mechanical energy E_f when the alpha-particle momentarily stops. The initial energy E_i is just the kinetic energy K_α of the incoming alpha-particle. The final energy E_f is just the electric potential energy U of the system. The potential energy U can be calculated from Eq. (13.1).

Let d be the centre-to-centre distance between the alpha-particle and the gold nucleus when the alpha-particle is at its stopping point. Then we can write the conservation of energy $E_i = E_f$ as

$$K_\alpha = \frac{1}{4\pi\epsilon_0} \frac{(2e)(Ze)}{d},$$

in which $Z = 79$ and $K_\alpha = 5.5$ MeV. Therefore,

$$d = \frac{1}{4\pi\epsilon_0} \frac{(2e)(79e)}{K_\alpha}$$

Now,

$$(1/4\pi\epsilon_0) = 9.0 \times 10^9 \text{ N m}^2 \text{ C}^{-2}$$

$$e = 1.6 \times 10^{-19} \text{ C}$$

$$\text{and } K_\alpha = 5.5 \text{ MeV} = 8.8 \times 10^{-13} \text{ J}$$

We then have,

$$\begin{aligned} d &= \frac{9.0 \times 10^9 \times 2 \times 79 \times (1.6 \times 10^{-19})^2}{8.8 \times 10^{-13}} \\ &= 4.13 \times 10^{-14} \text{ m} \\ &= 41.3 \text{ fm} \end{aligned}$$

This is a small distance on the atomic scale, well under 10^{-9} m (atomic dimensions), but not on the nuclear scale. It is, in fact, considerably larger than the sum of the radii of the gold nucleus and the alpha-particle. Thus, the alpha-particle reverses its motion without ever actually 'touching' the gold nucleus.

Thus, the experiments carried out by Geiger and Marsden gave rise to the Rutherford's nuclear model of the atom. According to this model, the positive charge of the atom is concentrated in a very small volume of the atom called nucleus. The nuclear radius is about $(1/10,000)$ of the atomic radius. To account for the fact that the electrons in an atom remain at relatively large distances from the nucleus, in spite of the electrostatic force of attraction of the nucleus for them, Rutherford postulated that the electrons *revolve* about the nucleus, the force of attraction providing the requisite centripetal force to keep them in their orbits.

13.3 ATOMIC SPECTRA

It is mentioned in section 13.1 that each element has a characteristic spectrum of radiation, which it emits. The spectrum consists of a set of isolated parallel lines. A spectrum of this kind is termed as a *line spectrum*. The spectrum emitted by atomic hydrogen is shown in Fig. 13.5. The wavelengths of the lines are characteristic of the element emitting the radiation.

We might expect that the frequencies of the light emitted by a particular element would exhibit some regular pattern. For instance, a radiating atom of an element, like a vibrating string, could emit a fundamental and its

harmonics. In the observed spectrum, however, at first sight, there does not seem to be any semblance of order or regularity. For many years unsuccessful attempts were made to correlate the observed frequencies with those of a fundamental and its overtones. Finally, in 1885, Johann Jakob Balmer (1825-1898) found a simple formula, which gave the frequencies of a group of lines emitted by atomic hydrogen.

Under proper conditions of excitation, atomic hydrogen emits the sequence of lines shown in Fig. 13.5. This sequence is called a series. There is evidently a certain order in this spectrum, the lines becoming crowded more and more closely as the limit of the series is approached. The line of the longest wavelength or the lowest frequency, in the red, is known as H_α , the next, in the blue-green, as H_β , the third as H_γ , and so on. Balmer found that the wavelengths of these lines were given by the simple formula

$$\frac{1}{\lambda} = R \left(\frac{1}{2^2} - \frac{1}{n^2} \right) \quad (13.3)$$

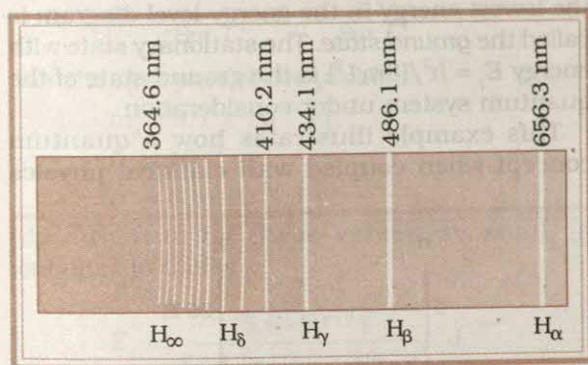


Fig. 13.5 Emission lines in the spectrum of atomic hydrogen.

where λ is the wavelength, R is a constant called the Rydberg constant, and n may have integral values 3, 4, 5, etc. The value of R is $1.097 \times 10^7 \text{ m}^{-1}$.

Letting $n = 3$ in Eq. (13.3), one obtains the wavelength of the H_α -line:

$$\begin{aligned} \frac{1}{\lambda} &= 1.097 \times 10^7 \text{ m}^{-1} (1/4 - 1/9) \\ \lambda &= 1.522 \times 10^6 \text{ m}^{-1} \end{aligned}$$

i.e., $\lambda = 656.3 \text{ nm}$

For $n = 4$, one obtains the wavelength of H_β -line, etc. For $n = \infty$, one obtains the limit of the series, at $\lambda = 364.6 \text{ nm}$. This is the shortest wavelength in the series.

Other series of spectra for hydrogen have since been discovered. These are known, after their discoverers, as Lyman, Paschen, Brackett, and Pfund series. These are represented by the formulae:

Lyman series:

$$\frac{1}{\lambda} = R \left(\frac{1}{1^2} - \frac{1}{n^2} \right) \quad n = 2, 3, 4, \dots$$

Paschen series:

$$\frac{1}{\lambda} = R \left(\frac{1}{3^2} - \frac{1}{n^2} \right) \quad n = 4, 5, \dots$$

Brackett series:

$$\frac{1}{\lambda} = R \left(\frac{1}{4^2} - \frac{1}{n^2} \right) \quad n = 5, 6, \dots$$

Pfund series:

$$\frac{1}{\lambda} = R \left(\frac{1}{5^2} - \frac{1}{n^2} \right) \quad n = 6, 7, 8, \dots$$

The Lyman series is in the ultraviolet, and the Paschen and Brackett series are in the infrared.

The Balmer formula Eq. (13.3) may be written in terms of frequency of the light, recalling that $c = v\lambda$.

$$\text{or} \quad \frac{1}{\lambda} = \frac{v}{c}$$

Thus, Eq. (13.3) becomes

$$v = Rc \left(\frac{1}{2^2} - \frac{1}{n^2} \right) \quad (13.4)$$

There are only a few elements (hydrogen, singly ionised helium, and doubly ionised lithium) whose spectra can be represented by simple formula like Eq. (13.4). Nevertheless, it is possible to separate the more complicated spectra of other elements into series, and to express the frequency of each line in the series as the **difference of two terms**.

13.4 ENERGY QUANTISATION

In your previous class you have learnt that a stretched string, of length L , clamped at both the ends can oscillate only in fixed modes for which the wavelengths are given by the set

$$\Lambda = \left\{ \lambda = \frac{2L}{n}; n = 1, 2, 3, \dots \right\} \quad (13.5a)$$

The frequencies of these modes are given by

$$\begin{aligned} v &= \frac{v}{\lambda} \\ &= n \frac{v}{2L} ; n = 1, 2, 3... \end{aligned} \quad (13.5b)$$

where v is the speed of travelling wave on the string (Section 15.7.1, Class XI). Thus, the vibrations of a bounded string provide an interesting example of discretisation or quantisation in classical physics.

In quantum mechanics, the analogue of a vibrating string is provided by a free particle of mass m confined to a line of length L . In Chapter 12, you have learnt that particles have wavelike properties with wavelength given by the de Broglie relation

$$p = \frac{h}{\lambda} \quad (13.6)$$

where p is the momentum of the particle. Pursuing the analogy between the string and the particle, the particle can have only those values of momentum for which the de Broglie wavelengths belong to the set Λ given by Eq. 13.5(a). Using Eqs. 13.5(a) and (13.6) we have

$$\begin{aligned} p &= \frac{h}{(2L/n)} \\ &= \frac{nh}{2L} ; n = 1, 2, 3... \end{aligned} \quad (13.7)$$

We see from Eq. (13.7) that the particle can only have discrete momenta or the magnitude of its momentum is quantised.

The energy of a free particle in terms of its momentum is given by

$$E = \frac{p^2}{2m} \quad (13.8a)$$

Substituting for p from Eq. (13.7) gives a simple result of quantum mechanics, namely that the energy of a free particle confined to a line of length L can have only discrete values of energy given by the relation

$$E_n = \frac{n^2 h^2}{8m L^2} ; n = 1, 2, 3... \quad (13.8b)$$

The parameter n , which takes the integer values,

1, 2, 3..., labels the stationary states* in ascending order of energy. Therefore, the energy of a free particle confined to a line of length L cannot be any real number from zero to infinity (as expected from classical mechanics); it can be only one of the following set of values:

$$E_1 = \frac{h^2}{8m L^2}, E_2 = \frac{4h^2}{8m L^2}$$

$$E_3 = \frac{9h^2}{8m L^2}, E_4 = \frac{16h^2}{8m L^2} \dots$$

The integer n , specifying the quantised energy states of the system, is called a *quantum number*.

Energy level diagrams usually depict the stationary states of a quantum system. In these diagrams a horizontal line represents the energy of a stationary state. Its height, from a reference level, is proportional to the energy of the stationary state. In Fig. 13.6, we show the first five stationary states of a particle moving freely along a line of length L . The stationary state of the lowest energy in the energy level diagram is called the *ground state*. The stationary state with energy $E_1 = h^2/(8mL^2)$ is the ground state of the quantum system under consideration.

This example illustrates how a quantum concept when coupled with classical physics

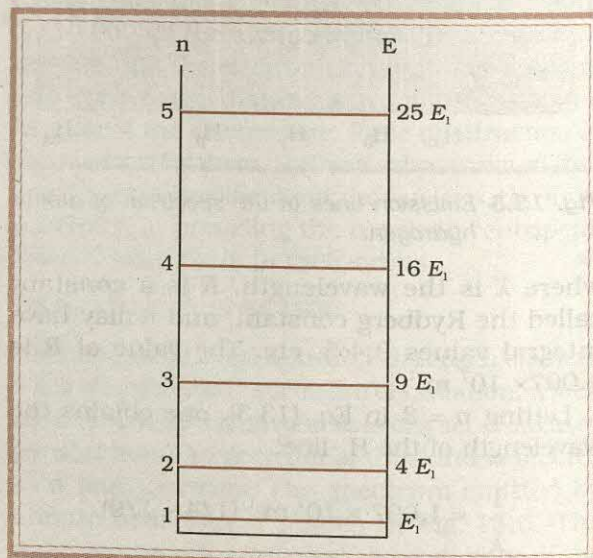


Fig. 13.6 Energy level diagram showing the first five stationary states of a particle of mass m moving freely along a line of length L .

* The word 'stationary state' appears in this section without any explanation. Its meaning will appear in Bohr's model.

introduces discreteness in the allowed energy levels. We shall make use of a similar argument in Bohr's model of the hydrogen atom to understand the spectrum of atomic hydrogen.

Example 13.2 Calculate the energies, in units of eV, for the ground state and the first excited state of an electron confined to a line of length 10^{-10} m (atomic dimensions) and also find the energies, in units of MeV, for the ground state and the first excited state of a neutron confined to a line of length 10^{-14} m (nuclear dimensions).

Answer The energies of the ground state and the first excited state of a free particle confined to a line correspond to $n = 1$ and $n = 2$ in the formula

$$E_n = \frac{n^2 h^2}{8m L^2}$$

$$E_1 = \frac{h^2}{8m L^2}, E_2 = \frac{4h^2}{8m L^2} = 4 E_1$$

For the first part of the problem

$$m_e = 9.1 \times 10^{-31} \text{ kg}$$

$$L = 10^{-10} \text{ m}$$

$$h = 6.6 \times 10^{-34} \text{ J s}$$

By substituting these values we shall get energies in joules.

$$E_1 = \frac{(6.6 \times 10^{-34})^2}{8 \times 9.1 \times 10^{-31} \times (10^{-10})^2} \text{ J}$$

$$= 5.98 \times 10^{-18} \text{ J}$$

$$1 \text{ eV} = 1.6 \times 10^{-19} \text{ J}$$

$$E_1 = (5.98 \times 10^{-18}) / (1.6 \times 10^{-19})$$

$$= 37 \text{ eV}$$

$$E_2 = 4 E_1$$

$$= 148 \text{ eV}$$

For the second part of the problem

$$m_n = 1.67 \times 10^{-27} \text{ kg}$$

$$L = 10^{-14} \text{ m}$$

$$E_1 = \frac{(6.6 \times 10^{-34})^2}{8 \times 1.67 \times 10^{-27} \times (10^{-14})^2} \text{ J}$$

$$= 3.26 \times 10^{-13} \text{ J}$$

$$1 \text{ MeV} = 1.6 \times 10^{-13} \text{ J}$$

$$E_1 = 2.0 \text{ MeV}$$

$$E_2 = 8.0 \text{ MeV}$$

13.5 BOHR'S ATOMIC MODEL AND THE HYDROGEN SPECTRUM

Rutherford nuclear model of the atom, discussed earlier has an unfortunate consequence. An electron moving in a circle is continuously accelerated toward the centre of the circle. According to classical electromagnetic theory (Chapter 9), an accelerated electron radiates energy. The total energy of an orbiting electron should therefore continuously decrease; its orbits become smaller and smaller, and eventually it should spiral into the nucleus and come to rest. But this goes against the observed stability of atoms. Furthermore, according to the classical theory, the frequency of the electromagnetic waves emitted by the revolving electrons is equal to the frequency of revolution. As the electrons radiate energy, their angular velocities would change continuously and they would emit a continuous spectrum, in contradiction to the line spectrum actually observed.

Faced with the above dilemma, in 1913, Niels Bohr concluded that in spite of the success of electromagnetic theory in explaining large-scale phenomena, it could not be applied to processes at the atomic scale. It became clear that a fairly radical departure from the established principles of classical mechanics and electromagnetism would be needed to understand the structure of atoms and the relation of atomic structure to atomic spectra.

Niels Bohr resolved the problem of the Rutherford's model by cleverly combining classical and early quantum concepts in the form of three postulates. Bohr's first postulate was that **an electron in an atom could revolve in certain stable orbits without the emission of radiant energy**, contrary to the predictions of electromagnetic theory. According to this postulate, each atom has certain definite stable states in which it can exist, and each possible state has definite total energy. These are called the stationary states of the atom. Bohr's second postulate defines these stable orbits. This postulate states that the electron revolves around the nucleus **only in those orbits for which the angular momentum is some integral multiple of $h/2\pi$** .

A completely stable atom, however, is as unsatisfactory as an unstable one, since atoms do emit radiant energy. Bohr's third postulate incorporated into atomic theory the early quantum concepts that had been developed by Planck and Einstein. His third postulate was that *an electron might make a transition from one of its specified non-radiating orbits to another of lower energy. When it does so, a single photon is emitted having energy equal to the energy difference between the initial and final states. The frequency of the emitted photon is then given by*

$$h\nu = E_i - E_f \quad (13.9)$$

where h is the Planck's constant and E_i and E_f are the energies of the initial and final states.

Each of these assumptions – the quantisation condition, the lack of radiation while in one of the quantised orbits, and the radiation in leaps between the orbits – was contrary to what was then known of classical theory. However, it was necessary to postulate the stability of the atom in some way. The radiation in jumps seemed to be consistent with what had already been revealed by Einstein and Planck. And the quantisation condition was not too different from the original condition of Planck. Let us now follow this mixture of classical and non-classical postulates to see the kind of atom that Bohr obtained.

The third postulate, if correct, already sheds some light on the hydrogen spectra described in Section 13.3. The Eq. (13.4) for the Balmer series in the hydrogen spectrum can also be written as:

$$h\nu = \frac{Rch}{2^2} - \frac{Rch}{n^2} \quad (13.10)$$

If we now identify the term $-Rch/n^2$ with the initial energy of the atom E_i , and $-Rch/2^2$ with its final energy, E_f , the Eq. (13.10) takes on the same form as Eq. (13.9). More generally, if we assume that the possible energy levels for a hydrogen atom are given by

$$E_n = -\frac{Rch}{n^2}; n = 1, 2, 3, \dots \quad (13.11)$$

then all the series spectra of hydrogen can be understood on the basis of transitions from one energy level to another. For the Lyman series the final state always corresponds to $n = 1$, for Paschen series it is $n = 3$ and so on. Similarly, complex spectra of other elements, represented

by two terms, are understood on the basis that each term corresponds to an energy level; and a frequency, represented as a difference of two terms, corresponds to a transition between these two energy levels.

The Eq. (13.11) has been guessed by combining the third postulate with the empirical formulae for Balmer and other spectral series for hydrogen. We now show that it can be derived from the second postulate and the usual classical mechanics.

Hydrogen is the simplest atom consisting of one electron orbiting around the nucleus having one proton. Bohr developed a model, which predicted correctly the energy levels of hydrogen atom. He assumed that the electron in the hydrogen atom moves in circular orbits, though orbits under inverse square force are, in general, elliptical. Planets move in elliptical orbits under the inverse square gravitational force of the sun. We now know that the assumption of circular orbits is not essential in a proper quantum mechanical treatment.

For an electron moving with a uniform speed in a circular orbit of a given radius, the centripetal acceleration is provided by the inverse square Coulomb force of attraction between the electron and the nucleus, the proton. The electron and proton each carry one unit of fundamental charge, e , but have opposite sign. The gravitational attraction between them is weaker than the Coulomb attraction by a factor of 10^{-40} and its contribution to the centripetal force can be neglected.

Let m_e be the mass of the electron. As the nucleus is about 2000 times heavier than the electron, it is reasonable to assume that it remains at rest. The centripetal acceleration for a particle moving in a circular orbit of radius r with speed v is v^2/r . Therefore, Newton's second law of motion gives

$$\frac{mv^2}{r} = \frac{e^2}{4\pi\epsilon_0 r^2} \quad (13.12)$$

giving

$$r = \frac{e^2}{4\pi\epsilon_0 mv^2} \quad (13.13)$$

The angular momentum of an electron moving in a circular orbit is given by

$$L = mvr$$

Now, the Bohr's second postulate leads to the condition

$$m v r = \frac{n h}{2 \pi} \quad (13.14)$$

where n is an integer and takes values 1, 2, 3,...

One can arrive at Eq. (13.14) in different manner. A circular orbit is taken to be a stationary orbit if its circumference contains integral number of de Broglie wavelengths. This is graphically depicted in Fig. 13.7 and can be mathematically expressed as

$$2 \pi r = n \lambda = n \left(\frac{h}{m v} \right) \quad (13.14a)$$

where n is an integer and takes values 1, 2, 3... This relation leads to the condition expressed by Eq. (13.14).

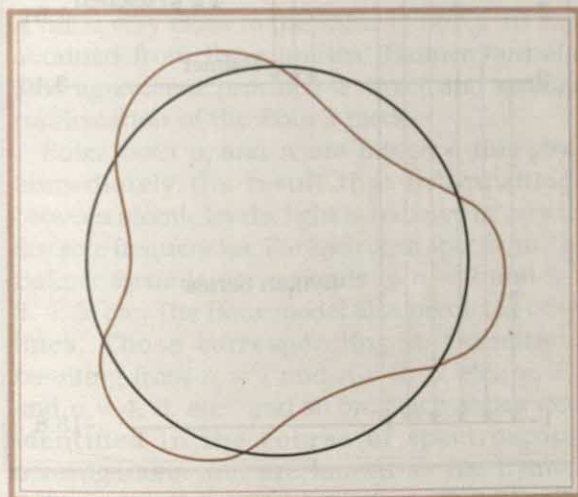


Fig. 13.7 The orbits satisfying the conditions of stationary states with three waves. The radius of the orbit with n waves can be estimated from Eq. (13.17).

Equation (13.14) is the Bohr's quantisation condition – one of his most radical postulates. As we shall see, the quantum condition [Eq. (13.14)] restricts the electron to certain possible orbits and yields the energy expression similar to as given in Eq. (13.11). However, in each of these orbits it should still, according to Maxwell's theory, radiate energy. Bohr solved this problem by fiat (if one can use that word). He asserted that, contrary to Maxwell's theory, contrary to Hertz's experiments, contrary to what everyone had believed until that time, electron does not radiate electromagnetic energy when it

is in a stationary orbit – an orbit allowed by the quantum condition [Eq. (13.14)]. When does then it radiate energy? This was the origin of Bohr's third postulate. Only when the electron makes a transition from one allowed orbit to another, Bohr asserted, does it radiate electromagnetic waves. And the precise frequency of radiation was to be given by Eq. (13.9).

The quantisation condition Eq. (13.14), gives a relation between the speed and radius of a possible orbit as

$$v = \frac{n h}{2 \pi m r} \quad (13.15)$$

Equations (13.13) and (13.15) can be solved for v and r . The speed of the electron and the radius of the orbit corresponding to a stationary state marked by the label n are given by the expressions:

$$v = \frac{1}{n} \left(\frac{e^2}{2 \epsilon_0 h} \right)$$

$$\text{or} \quad v = \frac{1}{n} \left(\frac{e^2}{4 \pi \epsilon_0} \right) \left(\frac{1}{h / 2 \pi} \right) \quad (13.16)$$

$$\text{and} \quad r = \left(\frac{n^2}{m} \right) \left(\frac{h}{2 \pi} \right)^2 \left(\frac{4 \pi \epsilon_0}{e^2} \right) \quad (13.17)$$

Eq. (13.16) depicts that the orbital speed in the outer orbits falls by the factor n . The size of the innermost orbit is called the *Bohr radius*, a_0 . It is given by

$$\begin{aligned} a_0 &= \frac{h^2 \epsilon_0}{\pi m e^2} \\ &= 5.29 \times 10^{-11} \text{ m} \end{aligned} \quad (13.18)$$

It can also be seen that the radii of the orbits increase as n^2 . The parameter n is called the *principal quantum number*.

The energy of the electron in the stationary states of the hydrogen atom can be obtained by adding its kinetic energy and the potential energy. The kinetic energy of the electron

$$= (1/2) m v^2$$

and its potential energy

$$= [(-e^2)/(4 \pi \epsilon_0 r)]$$

The total energy of the electron, E , in the atom is,

$$E = \frac{1}{2}mv^2 - \frac{e^2}{4\pi\epsilon_0 r} \quad (13.19)$$

Substituting the expressions for v and r in Eq. (13.19), we get the energies of the stationary states of the hydrogen atom,

$$E_n = -\frac{1}{2} \frac{mc^2}{n^2} \left(\frac{e^2}{4\pi\epsilon_0 (h/2\pi\kappa)} \right)^2$$

$$= -\frac{me^4}{8n^2\epsilon_0^2 h^2} \quad (13.20)$$

Equation (13.20) has exactly the same dependence on n as in Eq. (13.11) that we set out to prove.

When numerical values of the constants are inserted in Eq. (13.20) we obtain

$$E_n = -\frac{13.6}{n^2} \text{ eV} \quad (13.21)$$

The total energy has a negative sign because the zero of potential energy of the electron is at an infinite distance from the nucleus. The negative sign means that the electron is bound. Energy will thus be required to remove the electron from the hydrogen atom to a distance infinitely far away from its nucleus.

The derivation of Eq. (13.21) involves the assumption that the electronic orbits are circular. However, it was shown by Sommerfeld that, when the restriction of circular orbit is relaxed, Eq. (13.21) continues to hold even for elliptic orbits.

The energy of the atom is *least* (largest negative value) when its electron is revolving in an orbit closest to the nucleus i.e., the one for which $n = 1$. For $n = 2, 3, \dots$ the absolute value of the energy E is smaller, hence the energy is progressively larger in the outer orbits. The *normal* state of the atom, called the *ground state*, is that of the lowest energy, with the electron revolving in the orbit of smallest radius, the Bohr radius, a_0 . The energy of this state is -13.6 eV. Therefore, the minimum energy required to free the electron from the ground state of the hydrogen atom is 13.6 eV. It is called the *ionisation energy* of the hydrogen atom.

The hydrogen atom, normally, is in its ground state. When it receives energy by processes such

as electron collisions, the atom may temporarily acquire sufficient energy to raise the electron to higher orbits. The atom is then said to be in an excited state. From this state the electron can then fall back to a state of lower energy, emitting a photon in the process.

The energy level diagram for the stationary states of a hydrogen atom, computed from Eq. (13.21), is given in Fig. 13.8. The principal quantum number n labels the stationary states in the ascending order of energy.

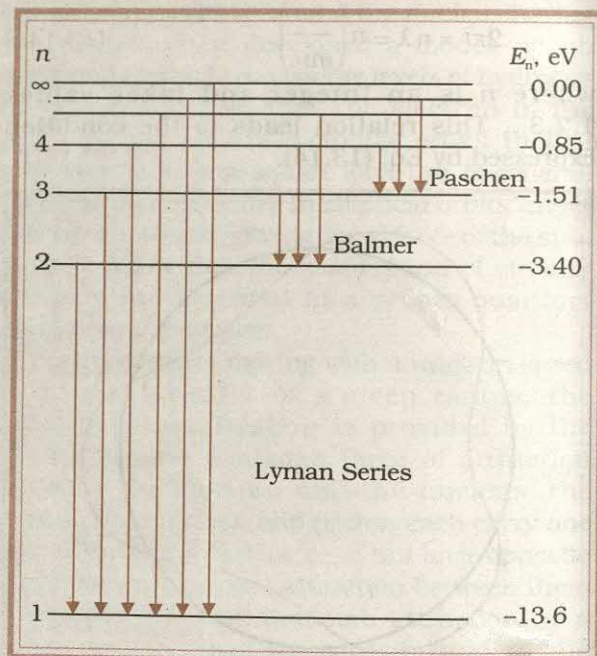


Fig. 13.8 The energy level diagram for the hydrogen atom. The energy level with $n = 1$ is the state of lowest energy and is called the ground state. A few of the transitions in the Lyman, Balmer and Paschen series are also shown.

Now, according to the third postulate, when an atom makes a transition from the state with quantum number n_i to the state with quantum number n_f ($n_f < n_i$), the difference of energy is carried away by a photon of frequency ν_f such that

$$h\nu_f = E_{n_i} - E_{n_f} \quad (13.22)$$

Using Eq. (13.20), for E_{n_f} and E_{n_i} , we get

$$h\nu_f = \frac{me^4}{8\epsilon_0^2 h^2} \left(\frac{1}{n_f^2} - \frac{1}{n_i^2} \right) \quad (13.23a)$$

$$\text{or } \nu_{if} = \frac{me^4}{8\epsilon_0^2 h^3} \left(\frac{1}{n_f^2} - \frac{1}{n_i^2} \right) \quad (13.23b)$$

Equation (13.23b) is the Rydberg formula, for the spectrum of the hydrogen atom. In this relation if we take $n_f = 2$ and $n_i = 3, 4, 5, \dots$, it reduces to a form similar to Eq. (13.4) for the Balmer series. The Rydberg constant R is readily identified to be

$$R = \frac{me^4}{8\epsilon_0^2 h^3 c} \quad (13.24)$$

If we insert the values of various constants in Eq. (13.24) we get

$$R = 1.03 \times 10^7 \text{ m}^{-1}$$

A value very close to the value ($1.097 \times 10^7 \text{ m}^{-1}$) obtained from the empirical Balmer formula. This agreement provided a direct and striking confirmation of the Bohr's model.

Since both n_f and n_i are integers, this gives immediately the result that in transitions between atomic levels, light is radiated in various discrete frequencies. For hydrogen spectrum, the Balmer formula corresponds to $n_f = 2$ and $n_i = 3, 4, 5$, etc. The Bohr model also predicted other lines. Those corresponding to transitions resulting from $n_f = 1$ and $n_i = 2, 3$, etc.; $n_f = 3$ and $n_i = 4, 5$, etc., and so on. Such series were identified in the course of spectroscopic investigations and are known as the Lyman, Balmer, Paschen, Brackett, and Pfund series. The electronic transitions corresponding to these series are shown in Fig. 13.9.

The explanation of the hydrogen atom spectrum provided by Bohr's model was brilliant achievement, which greatly stimulated progress toward the modern quantum theory. In 1922, Bohr was awarded Nobel Prize in Physics.

It may, however, be mentioned that Bohr's model of the hydrogen atom, involving classical trajectory picture (planet-like orbits around the nucleus), was inconsistent with the **Uncertainty Principle** and was replaced by the probabilistic picture of modern Quantum Mechanics. Interestingly, however, Bohr's formula for energy levels of a hydrogen atom [Eq. (13.20)] is exactly reproduced by quantum mechanics in non-relativistic domain.

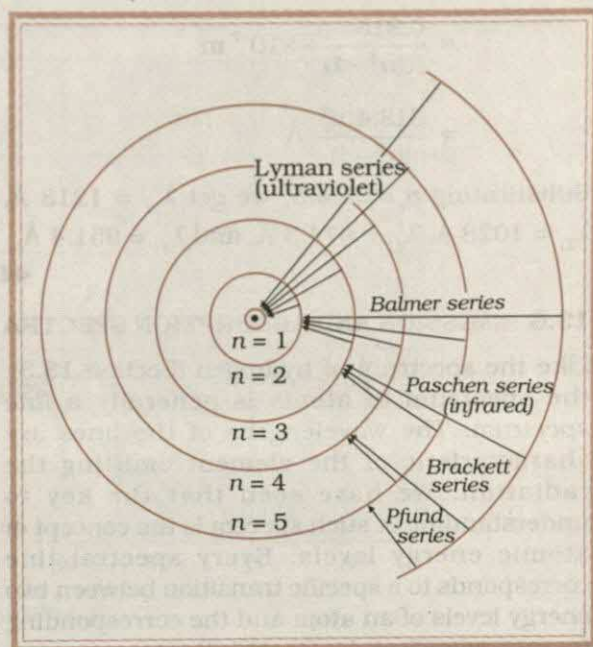


Fig. 13.9 'Permitted' orbits of an electron in the Bohr model of a hydrogen atom. The arrows indicate the transitions responsible for some of the lines of various series.

Example 13.3 Using the Rydberg formula, calculate the wavelengths of the first four spectral lines in the Lyman series of the hydrogen spectrum.

Answer The Rydberg formula is

$$hc/\lambda_{if} = \frac{me^4}{8\epsilon_0^2 h^2} \left(\frac{1}{n_f^2} - \frac{1}{n_i^2} \right)$$

The wavelengths of the first four lines in the Lyman series correspond to transitions from $n_i = 2, 3, 4, 5$ to $n_f = 1$. We know that

$$\frac{me^4}{8\epsilon_0^2 h^2} = 13.6 \text{ eV} = 21.76 \times 10^{-19} \text{ J}$$

Therefore,

$$\begin{aligned} \lambda_{if} &= \frac{hc}{21.76 \times 10^{-19} \left(\frac{1}{1} - \frac{1}{n_i^2} \right)} \text{ m} \\ &= \frac{6.625 \times 10^{-34} \times 3 \times 10^8 \times n_i^2}{21.76 \times 10^{-19} \times (n_i^2 - 1)} \text{ m} \end{aligned}$$

$$= \frac{0.9134 n_i^2}{(n_i^2 - 1)} \times 10^{-7} \text{ m}$$

$$= \frac{913.4 n_i^2}{(n_i^2 - 1)} \text{ \AA}$$

Substituting $n_i = 2, 3, 4, 5$, we get $\lambda_{21} = 1218 \text{ \AA}$, $\lambda_{31} = 1028 \text{ \AA}$, $\lambda_{41} = 974.3 \text{ \AA}$, and $\lambda_{51} = 951.4 \text{ \AA}$.

13.6 EMISSION AND ABSORPTION SPECTRA

Like the spectrum of hydrogen (Section 13.3), the spectrum of atoms is generally a *line spectrum*. The wavelengths of the lines are characteristic of the element emitting the radiation. We have seen that the key to understanding of such spectra is the concept of atomic energy levels. Every spectral line corresponds to a specific transition between two energy levels of an atom and the corresponding frequency is given by Eq. (13.9).

The lowest energy level of the atom is called the *ground state*, and all higher levels are called *excited states*. We have seen that a spectral line is emitted when an atom jumps or goes from an excited state to a lower state. Such a spectral line is called an *emission line*, and the lines corresponding to several such transitions constitute an *emission spectrum*.

The most familiar line spectrum is that of sodium. In the visible region, the sodium atom emits characteristic yellow light of wavelengths 589.0 and 589.6 nm when it undergoes transitions from the two excited levels to the ground state. Now, suppose the sodium atom is in the ground state and it were to *absorb* a quantum of radiant energy of wavelength 589.0 or 589.6 nm. It would then undergo a transition from the ground state to one of these levels. After a short time, the average value of which is called the *lifetime* of the excited state, the atom returns to the ground state and emits this quantum. For the excited levels of the sodium atom, the lifetime is about $1.6 \times 10^{-8} \text{ s}$.

A sodium atom in the ground state may absorb radiant energy of wavelengths other than the yellow lines. All wavelengths corresponding to spectral lines *emitted* when the sodium atom returns to normal ground state may also be *absorbed*. If, therefore, the continuous-spectrum light from a carbon arc is sent through an

absorption tube containing sodium vapours, and then examined with a spectroscope, there will be a series of dark lines corresponding to the wavelengths absorbed, as shown in Fig. 13.10. This is known as an *absorption spectrum*.

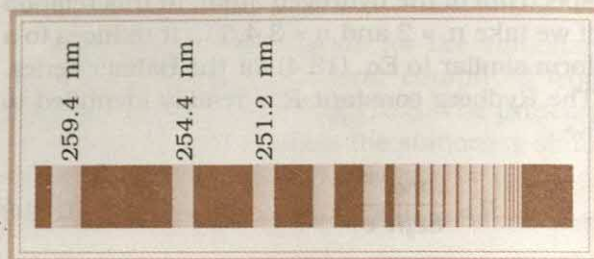


Fig. 13.10 Absorption spectrum of sodium.

The Fraunhofer lines seen in the sun's spectrum constitute an absorption spectrum. The main body of the sun emits continuous spectrum, whereas the cooler vapours in the sun's atmosphere emit line spectra corresponding to the elements present. When the intense light from the core of the sun passes through the cooler vapours in its atmosphere, the lines of these elements are absorbed. The light *emitted* by the cooler vapours is so faint as compared to unabsorbed continuous spectrum that the continuous spectrum appears to be crossed by many faint dark lines. These lines were first observed by Fraunhofer and are therefore called the *Fraunhofer lines*.

13.7 MANY ELECTRON ATOMS

The hydrogen atom discussed in Section 13.5 is the simplest of all atoms, containing one electron and one proton. The analysis of atoms with more than one electron was attempted on the lines of Bohr's model for hydrogen atom but did not meet with any success. Such an analysis increases in complexity very rapidly; each electron interacts not only with the positively charged nucleus but also with all the other electrons.

Several approximation schemes can be used; the simplest (and most drastic) is to ignore completely the interactions between electrons, and regard each electron as influenced by the electric charge of the nucleus, considered to be a point charge. A less drastic and more useful approximation is to think of all the electrons together as making up a charge cloud, that is on the average, *spherically symmetric*, and to

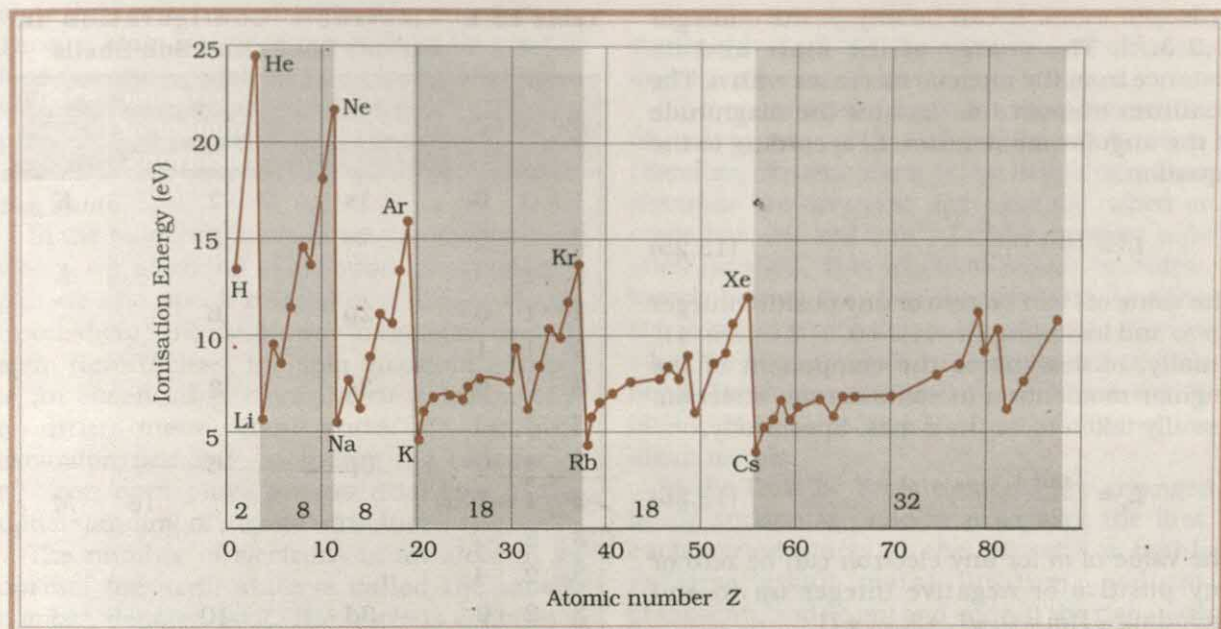


Fig. 13.11 A plot of the ionisation energies of the elements as a function of atomic number, the periodic repetition through the six complete periods of the Periodic Table. The number of elements in each of these periods is indicated.

think of each individual electron as moving in the total electric field due to the nucleus and this averaged-out electron cloud. This is known as the central field approximation; it provides a useful starting point for the understanding of atomic structure.

The central-field model suggests that each electron has a lowest energy state (roughly corresponding to the $n = 1$ state for a hydrogen-like atom with one nuclear charge). It might be expected that, in the ground state of a complex atom, all the electrons should be in the lowest state. If this is the case then there would be a gradual change in the physical and chemical properties of elements with increasing number of electrons.

A variety of evidence shows that this is *not* what happens at all.

The number of electrons per atom in the elements fluorine, neon, and sodium are respectively, 9, 10, and 11. Fluorine and neon are gases whereas sodium is a solid. Further, fluorine and sodium are chemically most active and neon is an inert gas.

The elements exhibit repetitive properties as function of their position in Periodic Table. The energy required to remove the most loosely bound electron from a neutral atom is called the

ionisation energy. The *ionisation potential* is the ionisation energy per unit charge. A plot of this energy of an element as a function of its position in the Periodic Table is shown in Fig. 13.11. It clearly exhibits a repetitive character. The remarkable similarity in the chemical and physical properties of the elements in each vertical column of the Periodic Table are evidence enough that the atoms are constructed according to systematic rules. In a complex atom all the electrons cannot be in the lowest energy state.

The key to this puzzle, discovered by the Swiss physicist Wolfgang Pauli in 1925, is called the *Pauli exclusion principle*. Briefly, it states that **no two electrons can occupy the same quantum-mechanical state**

To understand the application of the Pauli's principle to atomic structure, we need to use some results from quantum mechanics, the derivation of which is outside the scope of this book. First, the quantum-mechanical *state* of an electron in the central-field model is identified not by a single quantum number n as in Bohr model, but by a set of quantum numbers, all integers, usually called n , l , and m . The first number, n , is called the *principal quantum number*, corresponding to n for the

hydrogen atom. It can be any positive integer (1, 2, 3...). The energy of the state and its distance from the nucleus increases with n . The quantum number l designates the magnitude of the angular momentum, L , according to the equation,

$$L^2 = l(l+1) \left(\frac{h}{2\pi} \right)^2 \quad (13.25)$$

The value of l can be zero or any positive integer up to and including $(n-1)$ [$l = 0, 1, 2, \dots, (n-1)$]. Finally, m designates the component of the angular momentum in some chosen direction, usually taken to be the z -axis. Specifically,

$$L_z = \left(\frac{mh}{2\pi} \right) \quad (13.26)$$

The value of m for any electron can be zero or any positive or negative integer up to and including $\pm l$ [$m = 0, \pm 1, \pm 2, \dots, \pm l$].

These results imply quantised values for energy and angular momentum, just as the Bohr model did. It must be noted that in central field approximation, the energy level of an electron depends not only on n but also on l ; it does not depend on m . We can now make a list of the possible sets of quantum numbers and thus the possible energy levels of electrons in an atom. Such a list is given in Table 13.1. According to the notation used in this table, a level is designated by the quantum numbers n and l . A level for which $n = 2$ and $l = 0$ is called a 2s level and the state for which $n = 2$ and $l = 1$ is called a 2p level and so on. This table also shows the relation between values of n and the X-ray levels (K, L, M, \dots) described in the following section. Because the average electron distance from the nucleus increases with n , each value of n corresponds roughly to a region of space around the nucleus in the form of a spherical shell. Hence, one speaks of the L shell as that region occupied by the electrons in the $n = 2$ levels and so on. States with same n but different l are said to form *sub-shells*, such as the 3p sub-shell.

We are now better equipped for a precise statement of the Pauli's exclusion principle: **In any atom, not more than two electrons can occupy any quantum state.** That is, no more than two electrons can have the same set of three quantum numbers (n, l and m). It may not be clear why *two* electrons should be permitted in

Table 13.1 Electronic Configuration in various Shells and Sub-Shells.

n	l	m	Spectroscopic notation	Maximum Number of electrons		Shell
1	0	0	1s	2		K
2	0	0	2s	2		
2	1	-1	2p	6	8	L
2	1	0				
2	1	1				
3	0	0	3s	2		
3	1	-1	3p	6	18	M
3	1	0				
3	1	1				
3	2	-2	3d	10		
3	2	-1				
3	2	0				
3	2	1				
3	2	2				
4	0	0	4s	2		
4	1	-1	4p	6		
4	1	0				
4	1	1				
4	2	-2	4d	10	32	N
4	2	-1				
4	2	0				
4	2	1				
4	2	2	etc.			

each quantum state rather than just one. The reason is that we have not yet described the state completely. The missing component is the *electron spin*. In the Bohr model, electron spin is added by visualising the electron not as a point but as small ball of charge spinning on its axis. Associated with this spin is an angular momentum, and experiment shows that the component of this extra angular momentum in the direction of a specified axis (usually taken as the z -axis) is always one of the two values

$$\frac{1}{2} \left(\frac{h}{2\pi} \right) \text{ or } -\frac{1}{2} \left(\frac{h}{2\pi} \right).$$

Thus, for a given set of values of (n, l, m), there are two choices for the orientation of the spin angular momentum, corresponding to values ± 1 of a *fourth* quantum number s . When two electrons occupy the same

state they *must* have opposite spin orientations. Thus, a state is completely defined by a set of four quantum numbers (n , l , m , s). Now, when we include electron spin, the Pauli's exclusion principle can be restated: **In an atom, no two electrons can have all four quantum numbers the same.**

In the modern quantum-mechanical picture, where we abandon the notion of a localised particle and speak instead of a diffuse charge cloud, there is no simple way to visualise electron spin. Nevertheless, the spin quantum number is an essential part of the description of a quantum-mechanical state. The Pauli's exclusion principle including the concept of electron spin plays an essential role in the understanding of atomic structure.

The number of electrons in an atom in its normal (neutral) state is called the *atomic number*, denoted by Z . The nucleus contains Z protons and some neutrons. Because the electrons are attracted by the nucleus, we expect the quantum states corresponding to the regions near the nucleus to have lowest energy. We may imagine starting with a bare nucleus with Z protons, and adding electrons one by one until the normal complement of Z electrons for a neutral atom is reached. We expect the lowest-energy states, ordinarily those with smallest values of n and l , to fill first, and we use successively higher states until all the electrons are accommodated, thus filling various sub-shells and shells. The $1s$ state or the K shell accommodates two electrons. The $2s$ sub-shell

can again accommodate two electrons whereas the $2p$ sub-shell can accommodate six electrons and so on.

The outermost electrons determine the chemical properties of an element. It is, therefore, of particular interest to find how these electrons are arranged. For example, when an atom has one electron in the outermost sub-shell or shell, this electron would be rather loosely bound; the atom would tend to lose this electron and form with another suitable element what the chemists call an electrovalent or ionic bond. The valence of the element is then said to be $+1$. This behaviour is characteristic of the alkali metals.

In the Periodic Table elements are arranged in six horizontal **periods**; except for the first, each period starts at the left with a highly reactive alkali metal (lithium, sodium, potassium, and so on) and ends at the right with a chemically inert noble gas (neon, argon, krypton, and so on). Quantum physics accounts for the chemical properties of these elements. You would learn more about it in your chemistry course. The numbers of elements in the six periods are

2, 8, 8, 18, 18, and 32

Table 13.1 shows how these numbers arise naturally on the basis of the shell structure and Pauli's exclusion principle.

13.8 X-RAYS AND THE ATOMIC NUMBER

When a solid target, such as copper or tungsten is bombarded with electrons whose kinetic

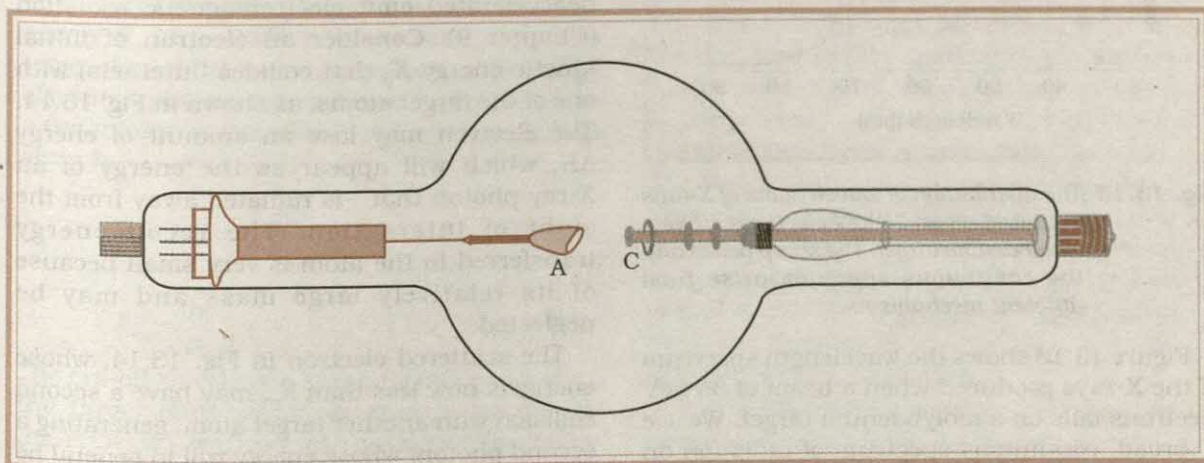


Fig. 13.12 A typical Coolidge type X-ray tube.

energies are in the kilo electron volt (keV) range, X-rays are emitted. They were first observed by Wilhelm K. Roentgen (1845-1923) in 1895 and were originally called *Roentgen Rays*.

X-rays are electromagnetic waves like light waves. The wavelength of X-rays ranges from 0.001 to 1 nm. The medical, dental and industrial usefulness of these is well known and widespread. Here our concern is to know what these rays can teach us about the atoms that absorb or emit them.

A common X-ray tube is the Coolidge type, invented by W.D. Coolidge of the General Electric Company in 1913. A typical Coolidge tube is shown in Fig. 13.12. It consists of a thermionic cathode C and an anode A made of copper or tungsten enclosed in a highly evacuated glass tube. Electrons are emitted from the cathode when the filament heats it. The electrons are accelerated through the potential difference between the two electrodes and strike the anode. X-rays are emitted from the anode surface as a consequence of its bombardment by the energetic electron stream.

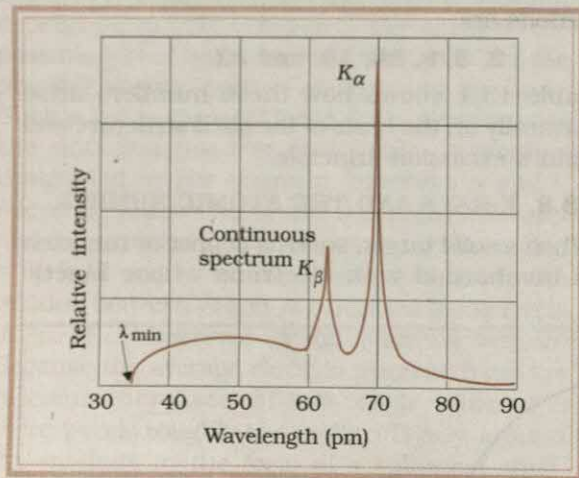


Fig. 13.13 The distribution of wavelengths of X-rays produced when 35 keV electrons strike a molybdenum target. The sharp peaks and the continuous spectrum arise from different mechanisms.

Figure 13.13 shows the wavelength spectrum of the X-rays produced when a beam of 35 keV electrons falls on a molybdenum target. We see a broad, continuous spectrum of radiation on which two peaks of sharply defined wavelengths are superimposed. The continuous spectrum

and the sharp peaks originate in different ways. We shall discuss this in the following sub-sections.

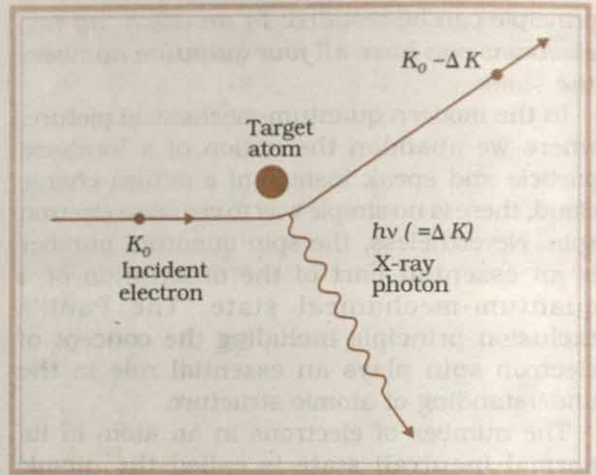


Fig. 13.14 An electron of kinetic energy K_0 passing near an atom in the target may generate an X-ray photon, the electron losing part of its energy in the process. The continuous spectrum arises in this way.

13.8.1 The Continuous X-ray Spectrum

Let us consider the X-ray spectrum shown in Fig. 13.13. It is a broad continuous spectrum with two sharp peaks. The X-rays constituting the continuous spectrum are sometimes also referred to as white X-rays. For the time being, we concentrate on the continuous part of the spectrum ignoring the two peaks. We recall that any charged particle accelerated or decelerated emit electromagnetic radiation (Chapter 9). Consider an electron of initial kinetic energy K_0 that collides (interacts) with one of the target atoms, as shown in Fig. 13.14. The electron may lose an amount of energy ΔK , which will appear as the energy of an X-ray photon that is radiated away from the sight of interaction. The recoil energy transferred to the atom is very small because of its relatively large mass and may be neglected.

The scattered electron in Fig. 13.14, whose energy is now less than K_0 , may have a second collision with another target atom, generating a second photon, whose energy will in general be different from that of the photon produced in the first collision. The electron-scattering

process continues until the electron loses all its energy and comes to rest. All the photons generated by these collisions form part of the continuous X-ray spectrum.

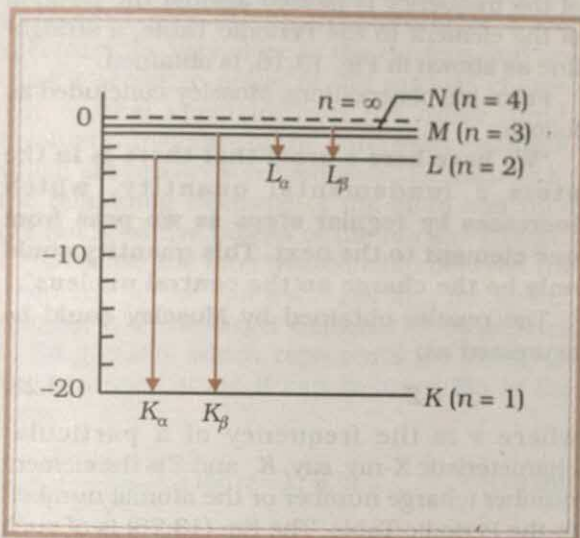


Fig. 13.15 An atomic energy level diagram for molybdenum, showing the transitions that represent the characteristic x-ray spectrum of that element. All levels except the K level consist of a number of close-lying components, which are not shown in the figure.

A prominent feature of the spectrum shown in Fig. 13.13 is the sharply defined **cut-off wavelength** λ_{\min} below which the continuous spectrum does not exist. This minimum wavelength corresponds to a collision in which an incident electron loses all its initial kinetic energy K_0 in a single head-on collision with a target atom. All this energy appears as the energy of a single photon, whose associated wavelength – the minimum possible X-ray wavelength – is given by

$$K_0 = h\nu = \left(\frac{hc}{\lambda_{\min}} \right) \quad (13.27a)$$

$$\text{or} \quad \lambda_{\min} = \frac{hc}{K_0} \quad (13.27b)$$

The cut-off wavelength is totally independent of the target material. If we change the target material, all features of the X-ray spectrum shown in Fig. 13.13 would change except the

cut-off wavelength, which depends only on the initial energy of the electrons hitting the target.

Example 13.4 A beam of 35.0 keV electrons strikes a molybdenum target. What is the cut-off wavelength of the X-rays generated?

Answer From the Eq. (13.27b) we have

$$\begin{aligned} \lambda_{\min} &= \frac{(4.14 \times 10^{-15} \text{ eV}\cdot\text{s}) \times (3.0 \times 10^8 \text{ m s}^{-1})}{35.0 \times 10^3 \text{ eV}} \\ &= 3.55 \times 10^{-11} \text{ m} \end{aligned}$$

White X-rays are employed for medical diagnostics. White X-rays with short λ_{\min} are termed as hard X-rays and employed for examining denser structures like bones. The white X-rays having longer λ_{\min} are termed as soft X-rays and used for examining softer tissues.

13.8.2 The Characteristic X-Rays

We now turn our attention to the two sharp peaks labelled as K_α and K_β in the X-ray spectrum shown in Fig. 13.13 (and other peaks that appear at wavelengths beyond the wavelength range displayed in this figure). The sharp peaks such as K_α and K_β form the characteristic spectrum of the target material.

The peaks arise in a two-step process: (1) An energetic electron strikes an atom in the target, while it is being scattered, the incident electron knocks out one of the atom's deep-lying (low n value) electrons. If the deep-lying electron is in the shell defined by $n = 1$ (the K shell), there remains a vacancy or *hole* in this shell. (2) An electron in one of the shells with a higher energy jumps to the K shell, filling the hole in this shell. During this process the atom emits a characteristic X-ray photon. If the electron that fills the K shell vacancy jumps from the shell with $n = 2$ (called the L shell), the emitted radiation is the K_α line of Fig. 13.13; if it jumps from the shell with $n = 3$ (called the M shell), it produces the K_β line and so on. An electron from still farther out in the atom will fill the hole left in the L or M shell.

A simplified atomic energy level diagram for molybdenum is shown in Fig. 13.15. The line at $E = 0$, represents the neutral atom in its

ground state. The level marked K (at $E = -20$ keV) represents the energy of the molybdenum atom with an electron in its K shell. Similarly, the level marked L (at $E = -2.7$ keV) represents the atom with an electron in the L shell and so on.

The transitions marked K_α and K_β in Fig. 13.15 are the ones that produce the X-ray peaks in Fig. 13.13. The K_α spectral line, as mentioned above originates when an electron from the L shell fills the hole in the K shell. In addition to the K series (K_α , K_β and K_γ) there are other series known as L , M , and N series, produced by the ejection of electrons from the L , M , and N shells rather than the K shell. As would be expected, the electrons in these outer shells, being farther away from the nucleus, are not held as firmly as those in the K shell. Consequently, more slowly moving electrons may excite the other series, and therefore, the photons emitted are of lower energy and longer wavelength. The characteristic X-rays constitute a line spectrum and are characteristic of the target element.

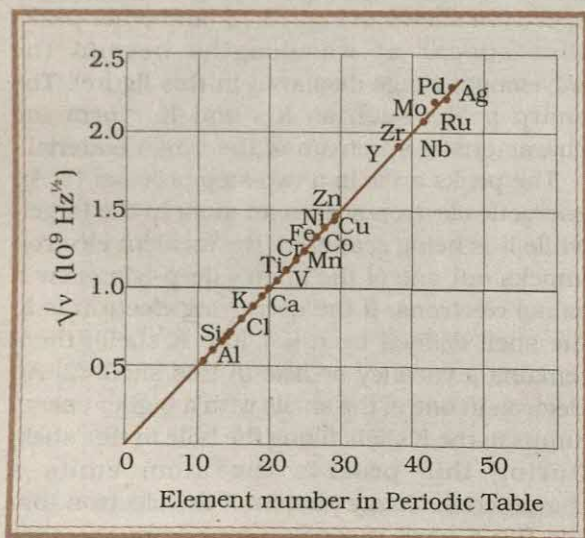


Fig. 13.16 A Moseley plot of the K_α line of the characteristic X-ray spectra of 21 elements. The frequency is calculated from the measured wavelength.

13.8.3 Numbering the Elements—Moseley's Law

In 1913, British physicist H.J.G. Moseley generated characteristic X-rays for as many as 38 elements by using them as targets for electron bombardment in an electron tube of his

own design. Moseley then sought for regularities in these spectra as he moved from element to element in the Periodic Table. He noted that if, for given spectral line, say, K_α , the square root of the frequency is plotted against the position of the element in the Periodic Table, a straight line as shown in Fig. 13.16, is obtained.

From his observations Moseley concluded as follows:

"We have here a proof that there is in the atom a fundamental quantity, which increases by regular steps as we pass from one element to the next. This quantity could only be the charge on the central nucleus".

The results obtained by Moseley could be expressed as:

$$\sqrt{\nu} \propto Z \quad (13.28)$$

where ν is the frequency of a particular characteristic X-ray, say, K_α and Z is the element number (charge number or the atomic number) in the Periodic Table. The Eq. (13.28) is of such a fundamental importance that it came to be known as Moseley's law.

Owing to Moseley's work, the characteristic X-ray spectrum became the universally accepted signature of an element, permitting a solution of number of Periodic Table puzzles. Prior to Moseley's work, the positions of elements in the Periodic Table were assigned in order of atomic mass, although it was necessary to invert this order for several pairs of elements because of compelling chemical evidence. Moseley showed that it is the nuclear charge (that is, the *atomic number* Z) and not the atomic mass, which is the real basis for numbering the elements.

In 1913 the Periodic Table had several empty places, and a surprising number of claims for new elements had been advanced. The X-ray spectrum provided a conclusive test of such claims. The lanthanide elements, often called the rare earth elements, had been sorted out only imperfectly because their similar chemical properties made sorting difficult. In more recent times, the identities of some elements beyond uranium were pinned down beyond any doubt only when the elements became available in quantities large enough to permit a study of their individual X-ray spectra.

It is not surprising to see why the characteristic X-ray spectrum shows such

impressive regularity from element to element whereas the optical spectrum in the visible or near visible region does not. The key to the identity of an element is the charge on the nucleus. The K electrons, which play an important role in the production of X-ray spectrum, lie very close to the nucleus and are thus a sensitive probe of the nuclear charge. The optical spectrum, on the other hand, involves transitions of the outermost electrons, which are heavily screened from the nucleus by the remaining electrons of the atom and thus are not sensitive probes of the nuclear charge.

To realise the inter-relationship between the frequency of the emitted X-ray and the atomic number Z of the target element, let us go back to Eq. (13.20), which represents the energy of the hydrogen atom. It can be rewritten in the form

$$E_n = -\frac{me^4}{8\epsilon_0^2 h^2} \frac{1}{n^2}; \text{ for } n = 1, 2, 3 \dots \quad (13.29)$$

Now, consider one of the two innermost electrons in the K shell of a multi-electron atom. Because of the presence of the other electron, the electron under consideration 'sees' an effective nuclear charge of approximately $(Z-1)e$, where e is the primary electronic charge and Z is the atomic number of the element. The factor e^4 in Eq. (13.29) is the product of $(+e)^2$ – the square of the charge of the hydrogen nucleus and $(-e)^2$ – the square of an electron's charge. For a multi-electron atom, we can approximate the effective energy of the atom by replacing e^4 in Eq. (13.29) with $(Z-1)^2 e^4$, that gives

$$E_n = -\frac{m(Z-1)^2 e^4}{8\epsilon_0^2 h^2} \frac{1}{n^2} \quad (13.30)$$

We saw that the K_α X-ray photon arises when an electron makes a transition from the L shell (with $n = 2$ and energy E_2) to the K shell (with $n = 1$ and energy E_1). Thus, using Eq. (13.30) we may write the energy change as

$$\begin{aligned} \Delta E &= -\frac{me^4}{8\epsilon_0^2 h^2} (Z-1)^2 \left(\frac{1}{2^2} - \frac{1}{1^2} \right) \\ &= -\frac{3me^4}{4 \times 8\epsilon_0^2 h^2} (Z-1)^2 \end{aligned} \quad (13.31)$$

Then the frequency of the K_α line is

$$\nu = \Delta E/h$$

$$= (2.46 \times 10^{15} \text{ Hz})(Z-1)^2$$

Taking square root of both the sides we have

$$\sqrt{\nu} = CZ - C \quad (13.32)$$

The Eq. (13.32) is equation of a straight line. It shows that if we plot the square root of the frequency of the K_α spectral line against the atomic number Z , we should get a straight line. This is exactly what Moseley found.

Example 13.5 A cobalt target is bombarded with electrons, and the wavelengths of its characteristic X-ray spectrum are measured. There is also a second, fainter characteristic spectrum, which is due to an impurity in cobalt. The wavelengths of K_α lines are 178.9 pm (cobalt) and 143.5 pm (impurity). Using this data identify the impurity.

Answer The wavelengths of the K_α lines for both the cobalt (Co) and the impurity (X) fall on a K_α Moseley plot, and Eq. (13.32) is the equation for the plot. Substituting c/λ for ν in that formula, we obtain

$$\sqrt{\frac{c}{\lambda_{\text{Co}}}} = CZ_{\text{Co}} - C \quad (i)$$

$$\text{and } \sqrt{\frac{c}{\lambda_X}} = CZ_X - C \quad (ii)$$

Dividing Eq.(i) by (ii) we get

$$\frac{\sqrt{\frac{c}{\lambda_X}}}{\sqrt{\frac{c}{\lambda_{\text{Co}}}}} = \frac{(Z_{\text{Co}} - 1)}{(Z_X - 1)}$$

Substituting the given data yields

$$\frac{\sqrt{143.5 \text{ pm}}}{\sqrt{178.9 \text{ pm}}} = \frac{27 - 1}{Z_X - 1}$$

Solving for the unknown, we get

$$Z_X = 30.0$$

A glance at the Periodic Table identifies the impurity as Zn.

13.9 SPONTANEOUS AND STIMULATED EMISSION – MASER AND LASER

In previous sections we have talked about the various energy levels of an atom. An isolated atom

can exist either in its state of lowest energy, the ground state whose energy is, say, E_0 , or in a state of higher energy (an excited state), whose energy is E_x . There are three processes by which the atom can move from one of these states to the other. These processes are shown in Fig. 13.17. We shall now discuss them one by one.

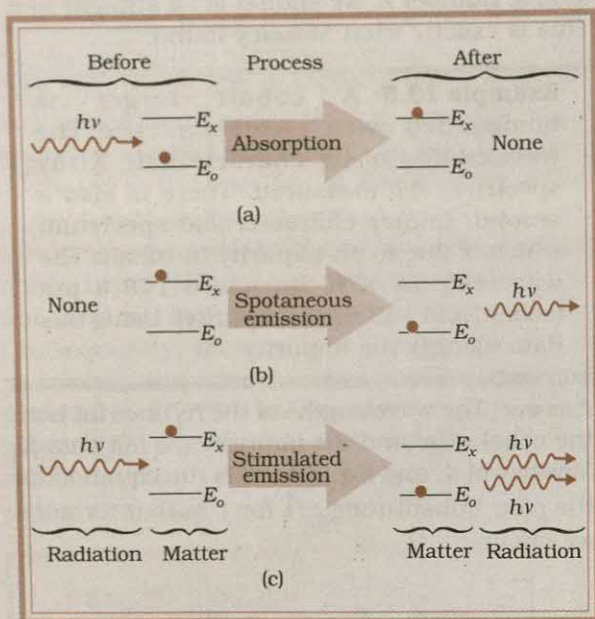


Fig. 13.17 The interaction of radiation and matter in the process of (a) absorption, (b) spontaneous emission, and (c) stimulated emission.

1. **Absorption:** Figure 13.17(a) shows that the atom is initially in its ground state. Now, if an electromagnetic radiation of frequency ν is incident on the atom, the atom can absorb an energy $h\nu$ (a photon) from the radiation and move to the higher energy state. From the principle of conservation of energy we have

$$h\nu = E_x - E_0 \quad (13.33)$$

We call this process **absorption**.

2. **Spontaneous emission:** Figure 13.17(b) shows that the atom is in its excited state and no external radiation is present. It comes to this state after absorbing a photon of energy $h\nu$ or by some other inelastic collision. The atom has now become an excited atom **A**. A short time later, the atom will move of its own accord to its ground state, emitting a photon of

energy $h\nu$. We call this process **spontaneous emission** – *spontaneous* because the event was not triggered by any outside influence. The direction and phase of each of such photons is random. The light from a sodium or mercury lamp is generated in this manner.

Normally, the mean-life of excited atoms before spontaneous emission takes place is about 10^{-8} s. However, for some of the excited states, this mean-life can be as much as 10^5 times longer. Such long-lived states are called *meta stable states*; as we shall see they play an important role in laser operation.

3. **Stimulated emission:** In Fig. 13.17(c) the atom is in its excited state but this time a radiation with a frequency given by the Eq. (13.33) is also simultaneously present. Under these circumstances a photon of energy $h\nu$ can stimulate the atom to move to its ground state, during which process the atom emits an additional photon, whose energy is also $h\nu$. We call this process **stimulated emission** – *stimulated* because the event is triggered by the external photon. The emitted photon is in every way identical to the stimulating photon. It has the same energy, phase, polarisation, and direction of travel.

As we shall see, the process of stimulated emission is the key to laser operation. Einstein introduced this concept in 1917. Although the world had to wait until 1960 to see an operating laser, the groundwork for its development was put in place decades earlier.

13.9.1 Laser

The word LASER is an acronym and stands for **light amplification by stimulated emission of radiation**, which sums up the operation of an important optical and electronic device. The laser is a source of highly directional, monochromatic, and coherent light.

The action of a laser is based on the principle of stimulated emission. Figure 13.17(c) describes stimulated emission for a single atom. Consider an absorption cell containing a large number of atoms of the type described by Fig. 13.17(c) in thermal equilibrium at temperature T . Before any radiation is directed at the sample, a number N_0 of these atoms are in the ground state with energy E_0 , and a number N_x are in a state of

higher energy E_x . Boltzmann showed that N_x is given in terms of N_0 by

$$N_x = N_0 e^{-(E_x - E_0)/kT} \quad (13.34)$$

where k is the Boltzmann constant. The quantity kT is the mean kinetic energy of the atom at temperature T . Thus, higher the temperature, more atoms will be thermally excited to the higher state. Also because $E_x > E_0$, Eq. (13.34) requires that $N_x < N_0$; that is, there will always be fewer atoms in the excited state than in the ground state. This is expected if the level populations N_0 and N_x are determined only by the action of thermal excitation. Such a situation is illustrated in Fig. 13.18(a).

If we now flood the atoms, represented in Fig. 13.18(a), by photons of energy $E_x - E_0$, photons will disappear via absorption by ground-state atoms, and will be generated largely via stimulated emission of the excited state. It can be shown that the probabilities per atom of these two processes are identical. Since there are more atoms in the ground state, the net effect will be the absorption of photons.

To have laser action, we must have more photons emitted than absorbed. Therefore, we must have a situation in which stimulated emission dominates. A direct way to bring this about is to start with more atoms in the excited state than in the ground state, as shown in Fig. 13.18(b). Such a situation is termed as **population inversion**. However, such a population inversion is not consistent with thermal equilibrium. Therefore, to create and maintain such a situation we have to think of some clever way.

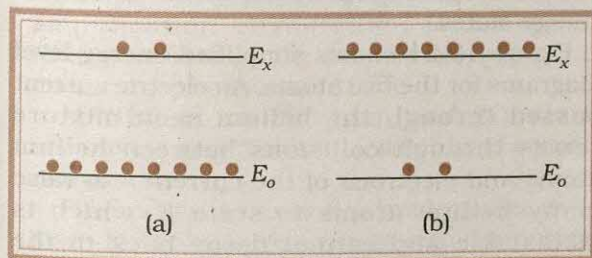


Fig. 13.18 (a) The equilibrium distribution of atoms between the ground state E_0 and the excited state E_x , accounted for by thermal agitation. (b) An inverted population, obtained by special methods. Such an inverted population is essential for laser action.

Let us now consider a system of atoms in which most of the atoms are initially in their ground state of energy E_0 . This atom has an excited level designated by energy E_3 , as shown in Fig. 13.19. The system is now irradiated by photons of energy $E_3 - E_0$ or excited by electron collisions in a discharge tube. As a result of this process, more and more atoms are excited to the state designated by the energy E_3 . This state is highly unstable and the atoms decay rapidly to a state designated by energy E_2 . The energy difference $E_3 - E_2$ is given up in the form of heat. Now, suppose the state E_2 has a long mean life (say, of the order of 5 ms). Such a state is called a *metastable state*. The level E_2 is very important for the stimulated emission process since the atoms in this state have a mean life of ~ 5 ms before they fall to the ground state. If the atoms are excited from E_0 to E_3 at a rate faster than the rate at which the atoms in state E_2 fall back to the ground state E_0 , the population of the metastable state E_2 becomes larger than that of the ground state E_0 . This is now a reversal of

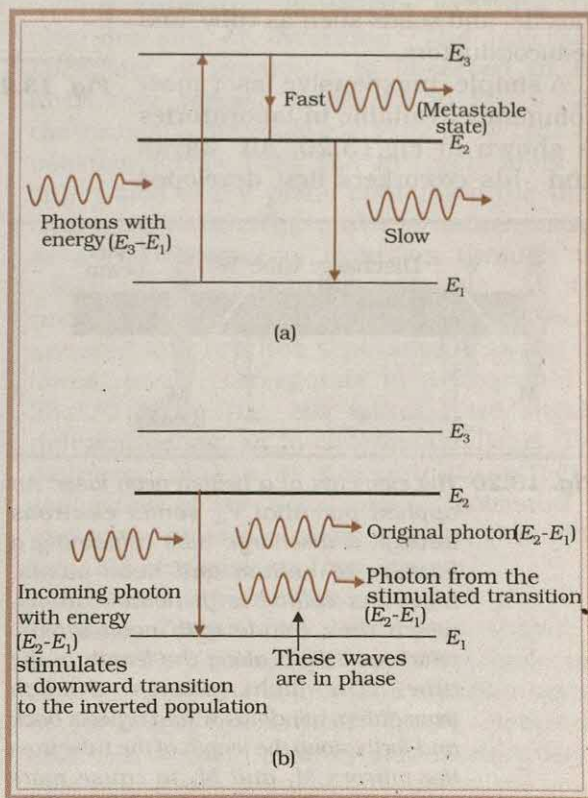


Fig. 13.19 (a) Generation of an inverted population, (b) stimulated emission.

the state of thermal equilibrium, where there are always fewer atoms in the higher energy states than the lower ones. Thus, a situation of *population inversion* has been created. The inverted population is crucial for laser action. Let us now consider what happens when a photon of energy $E_2 - E_0$ enters this system and interacts with one of the inverted population atoms. This photon can now actually stimulate the atom to fall from state E_2 to E_0 and emit a photon of energy $E_2 - E_0$. The first photon has stimulated the emission of another photon of same energy. This amounts to multiplying the number of photons in the system by a factor of two, and so on. We, therefore, have light amplification by stimulated emission of radiation or LASER.

Laser action has been obtained using many different materials, including gases, such as neon, carbon dioxide, and solids such as ruby, and semiconductors.

A simple, inexpensive laser most commonly available in laboratories is shown in Fig. 13.20. Ali Javan and his coworkers first developed

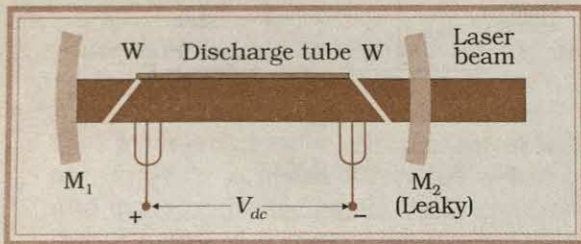


Fig. 13.20 The elements of a helium-neon laser. An applied potential V_{dc} sends electrons through a discharge tube containing a mixture of helium and neon gases. Electrons collide with helium atoms, which then collide with neon atoms, which emit light along the length of the tube. The light passes through transparent windows W and reflects back and forth along the length of the tube from the mirrors M_1 and M_2 to cause more emission from neon atoms. Some of the light leaks through the mirror M_2 to form the laser beam.

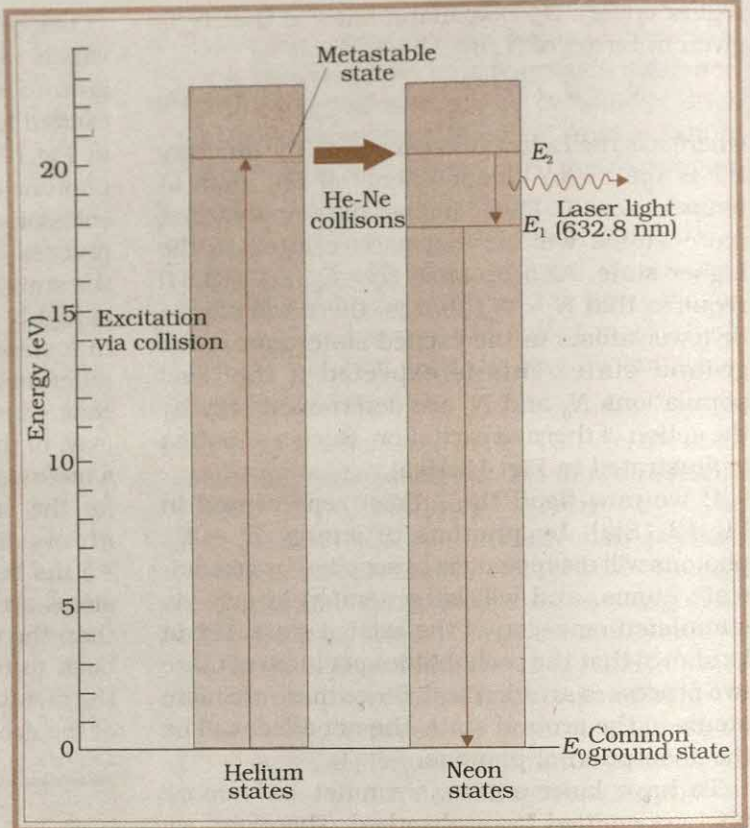


Fig. 13.21 Four essential energy levels in a helium and neon atoms in a helium-neon gas laser. Laser action occurs between the levels E_2 and E_1 of neon when more atoms are at the E_2 than at the E_1 level.

such a laser in 1961. A mixture of helium and neon in a proportion of 1:4 is filled in a glass discharge tube. When a sufficiently high voltage is applied, a glow discharge occurs. Collisions between ionised atoms and electrons carrying the discharge current excite atoms to various energy states.

Figure 13.21 shows simplified energy level diagrams for the two atoms. An electric current passed through the helium-neon mixture serves through collisions between helium atoms and electrons of the current – to raise many helium atoms to state E_3 , which is metastable and cannot decay back to the ground state.

The energy of the helium state E_3 (20.61 eV) is very close to the energy of neon state E_2 (20.66 eV). Thus, when a metastable (E_3) helium and a ground state (E_0) neon atom collide, the excitation energy of the helium atom is often transferred to the neon atom, which then moves

to state E_2 . In this manner neon level E_2 , shown in Fig. 13.21, can become more heavily populated than the neon level E_1 . Thus, we have the necessary mechanism for a population inversion in neon. This population inversion is relatively easy to set up because (1) initially there are essentially no neon atoms in the state E_1 , (2) the metastability of the helium level E_3 ensures a ready supply of neon atoms in level E_2 , and (3) atoms in level E_1 decay rapidly (through intermediate levels not shown) to the neon ground state E_0 .

Suppose now that a single photon is spontaneously emitted as a neon atom transfers from state E_2 to state E_1 . Such a photon can trigger stimulated emission event, which in turn can trigger other stimulated events. Through such a chain reaction, a coherent beam of red laser light at 632.8 nm, moving parallel to the tube axis is generated. In practice the beam is sent back and forth through the gas many times by a pair of mirrors M_1 and M_2 , so as to stimulate emission from as many excited atoms as possible. One of the mirrors M_1 is totally reflecting while the other one, M_2 , is partially leaky so that a small fraction of the laser light escapes to form a useful external beam.

The net effect of all the processes taking place in a laser tube is a beam of radiation, which has interesting properties described in subsection 13.9.3.

13.9.2 MASER

In section 13.9, it has been shown that in a system in which a situation of population inversion has been created, a photon of appropriate energy can induce stimulated emission of radiation. This process leads to amplification of radiation. If the emitted radiation falls in the microwave region, the device is termed as a MASER.

MASER is an acronym and stands for **microwave amplification by stimulated emission of radiation**.

Historically, a maser was made earlier than a laser. In 1953, J. Weber suggested that inverted populations could be used in an amplifying device, pointing out the possibility of pulse amplification from systems with inverted populations of energy level of magnetic ion. He also considered the use of a gas to obtain continuous operation. In 1954 and 1955, J.P. Gordon, H.J. Zeiger and C.H. Townes published papers describing the

continuous operation of an oscillating device working with an inverted population of two of the levels of ammonia molecule.

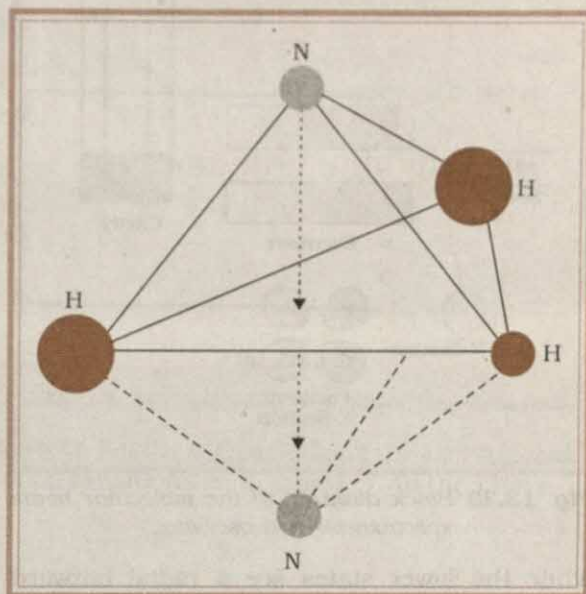


Fig. 13.22 Inversion of ammonia molecule.

The ammonia molecule exhibits the phenomenon of inversion. The inversion spectrum of ammonia is observed in the microwave region. As shown in Fig. 13.22, the ammonia molecule can exist in two equally possible forms with the nitrogen atom on either side of the plane containing the three hydrogen atoms. These two structures result as a consequence of inversion through the centre of mass of the molecule. All the rotational-vibrational levels of the molecule are split into two; the separation between the lowest levels corresponds to a frequency of 23,870 MHz. The two states have slightly different behaviour in electrostatic fields. This property makes it possible to effect a separation of the two types of molecules by passing a stream of ammonia gas through a suitable electrostatic lens system.

A block diagram of the set-up employed by Gordon, Zeiger and Townes is shown in Fig. 13.23. A beam of ammonia molecules emerges from the source and enters a system of focussing electrodes. These electrodes establish a quadrupolar cylindrical electrostatic field whose axis is in the direction of the beam. Of the inversion levels, the upper states experience a radial inward (focussing) force,

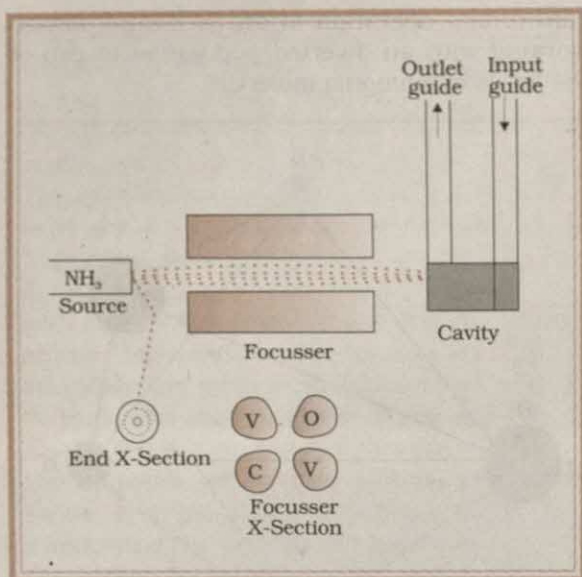


Fig. 13.23 Block diagram of the molecular beam spectrometer and oscillator.

while the lower states see a radial outward force. Therefore, the molecules in the lower states are drawn out of the beam resulting in a separation of the molecules in the two states. The molecules arriving at the cavity are then virtually all in the upper states. The cavity is tuned to the transition frequency between the two states. Transitions to the lower energy state are induced due to stimulation effects of the radiation present as noise in the cavity; these molecules emit radiation, which in turn cause more stimulated transitions. If the molecular beam is sufficiently intense, the cavity losses are overcome and the device oscillates continuously. The gas pressure has to be kept low in order to minimize collision processes by which the equilibrium population could be restored. Thus, the power output of such a device was limited to only about 10^{-8} W. The word Maser was first used in connection with this device.

After the discovery of the first maser, rapid developments took place. In place of a gas, solid materials were employed which provided higher power output and wider bandwidth. Masers were employed as first-stage amplifier in satellite communication systems. As most of the maser devices were operated at liquid helium temperatures and had smaller bandwidths < 100 MHz, their usefulness was

rather limited. These limitations of the masers, together with the very rapid developments, which have taken place in the area of diode parametric amplifiers, have resulted in a virtual cessation of development in this field.

13.9.3 Properties of Laser Light

(a) **Laser light is highly monochromatic:**

Light from an ordinary incandescent lamp is spread over a continuous range of wavelengths. The light from a fluorescent mercury lamp has several wavelengths; even light from a neon sign is monochromatic to only 1 part in about 10^6 . However, the frequency spread $\Delta\nu$ of laser light can be made really narrow with the

sharpness, $\frac{\Delta\nu}{\nu}$, many times greater, as much as 1 part in 10^{15} .

(b) **Laser light is highly coherent:** When two separated beams, originating from the same source, that have travelled long distances (over several hundred kilometres) over separate paths, are recombined, they 'remember' their common origin and are able to form a pattern of interference fringes. The corresponding coherence length for light originating from a light bulb is typically less than a metre.

(c) **Laser light is highly directional:** A laser beam spreads very little; it departs from parallelism only because of diffraction at the exit aperture of the laser. For example, a laser beam used to measure the distance to the Moon generates a spot on the Moon's surface with a diameter of only a few metres. The laser beam has an extremely small angular divergence.

(d) **Laser light can be sharply focussed:** If two light beams transport the same amount of energy, the beam that can be focussed to the smaller spot will have the greater intensity at that spot. For laser light the focussed spot can be so small that intensity of the order of 10^{17} W/cm² is readily obtained. An oxyacetylene flame has an intensity of only 10^3 W/cm².

13.9.4 Applications of Laser Light

In recent years, lasers have found a wide variety of applications. The high intensity of a laser

beam makes it a convenient drill. A very small hole can be drilled in a diamond for use as a die in drawing very small-diameter wires. The ability of a laser beam to travel long distances without appreciable divergence make it a very useful tool for surveyors, especially when great precision is required over long distances, as in the case of a long tunnel being drilled from both ends.

The smallest lasers, used for voice and data transmission over optical fibres, have as their active medium a semiconducting crystal about the size of a pinhead. Small as they are, they can generate about 200 mW of power. The largest lasers used for nuclear fusion research and for astronomical and military applications, fill a large building.

Among the many uses of lasers are reading bar codes, manufacturing and reading compact discs, cutting cloth in garment industry, welding auto bodies, etc.

Lasers are finding increasing applications in medical science. A laser can produce a very narrow beam with extremely high intensity, high enough to vaporize anything in its path. This property is used in the treatment of a detached retina; a short burst of radiation damages a small area of the retina, and the resulting scar tissue 'welds' the retina back to the choroids from which it has become detached. Laser beams are also used in surgery; blood vessels cut by laser beam tend to seal themselves off, making it easier to control bleeding.

Semiconductor lasers, as laser diodes, are finding wide application in optical communication.

Example 13.6 When sunlight shines on the atmosphere of Mars, carbon dioxide molecules at an altitude of about 75 km undergo natural laser action. The energy levels involved in the action are shown in the following figure; population inversion occurs between energy levels E_2 and E_1 . (a) What wavelength of sunlight excites the molecules in laser action? (b) At what wavelength does lasing occur? (c) In what region of the electromagnetic spectrum do the excitation and lasing wavelengths lie?

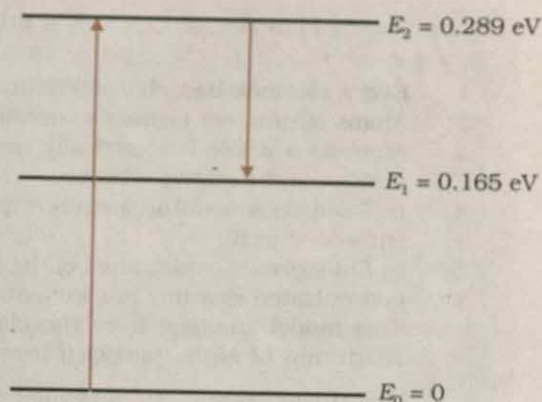


Fig. 13.24

Answer Let the wavelengths of the exciting and lasing radiations be λ_{ex} and λ_{La} . We then have

$$\frac{hc}{\lambda_{ex}} = E_2 - E_0$$

$$\begin{aligned} \text{or } \lambda_{ex} &= \frac{hc}{0.289 \text{ eV}} \\ &= \frac{6.625 \times 10^{-34} \text{ J} \times 3 \times 10^8 \text{ m}}{0.289 \times 1.6 \times 10^{-19} \text{ J}} \\ &= 42.98 \times 10^{-7} \text{ m} \\ &= 4298 \text{ \AA} \end{aligned}$$

Similarly,

$$\begin{aligned} \frac{hc}{\lambda_{La}} &= E_2 - E_1 \\ \lambda_{La} &= \frac{hc}{0.124 \text{ eV}} \\ &= \frac{6.625 \times 10^{-34} \text{ J} \times 3 \times 10^8 \text{ m}}{0.124 \times 1.6 \times 10^{-19} \text{ J}} \\ &= 100.17 \times 10^{-7} \text{ m} \\ &= 10017 \text{ \AA} \end{aligned}$$

Thus, the wavelength of sunlight, which excites lasing action, is 4298 Å. The lasing occurs at 10017 Å. The radiation, which excites lasing, lies at the violet end and the laser light lies at the infrared end of the electromagnetic spectrum.

SUMMARY

1. Every element has characteristic atoms.
 2. Atoms of different elements contain electrons, which are completely identical.
 3. Atom as a whole is electrically neutral and therefore contains equal amount of positive and negative charges.
 4. In Thomson's model an atom is a spherical cloud of positive charges with electrons embedded in it.
 5. In Rutherford's model, most of the mass of the atom and all its positive charge are concentrated in a tiny nucleus, and the electrons revolve around it.
- This model emerged from the classic experiments of Geiger and Marsden on scattering of alpha-particles from metal foils. A collimated beam of 5.5 MeV

α -particles from $^{214}_{83}\text{Bi}$ was allowed to fall on a $2.1 \times 10^{-7}\text{m}$ thin gold foil. The scattered α -particles produced scintillations on a ZnS screen, which were counted at different angles (θ) from the direction of the beam. Though most of the α -particles suffered negligible deviation, some suffered a large change in direction ($\theta > 90^\circ$). The last observation was a crucial clue to the nuclear model.

6. Rutherford's calculations used the inverse-square law of repulsive force between an α -particle ($Z = 2$) and a gold nucleus ($Z = 79$) ignoring multiple scattering. The scattering angle θ of the α -particle is related to the impact parameter b through the relation

$$b = \frac{Ze^2 \cot(\theta/2)}{4\pi\epsilon_0 E}$$

where E is the kinetic energy of the incident α -particle. The impact parameter b is the perpendicular distance of the initial velocity vector of the α -particle from the centre of the nucleus. The observed number of scattered α -particles at different angles agreed with Rutherford's calculations based on the nuclear model of an atom.

7. Classically, Rutherford's model of the atom is unstable. An orbiting electron accelerates continuously and must lose energy in electromagnetic radiation. The orbit must shrink spirally into the nucleus and give out a continuous spectrum of radiation.
8. Atoms of each element are stable and emit characteristic spectrum. The spectrum consists of a set of isolated parallel lines termed as line spectrum. It provides useful information about the atomic structure. The atomic hydrogen emits a line spectrum consisting of various series. The frequency of any line in a series can be expressed as a difference of two terms;

$$\text{Balmer series : } \nu = Rc \left(\frac{1}{2^2} - \frac{1}{n^2} \right); n = 3, 4, 5, \dots$$

$$\text{Lyman series: } \nu = Rc \left(\frac{1}{1^2} - \frac{1}{n^2} \right); n = 2, 3, 4, \dots$$

$$\text{Paschen series: } \nu = Rc \left(\frac{1}{3^2} - \frac{1}{n^2} \right); n = 4, 5, 6, \dots$$

$$\text{Brackett series: } \nu = Rc \left(\frac{1}{4^2} - \frac{1}{n^2} \right); n = 5, 6, 7, \dots$$

$$\text{Pfund series: } \nu = Rc \left(\frac{1}{5^2} - \frac{1}{n^2} \right); n = 6, 7, 8, \dots$$

9. In quantum mechanics, the energies of a system are discrete or quantised. The energy of a particle of mass m confined to a line of length L can have only discrete values of energy given by the relation

$$E_n = \frac{n^2 h^2}{8mL^2}; n = 1, 2, 3, \dots$$

10. Bohr's model of the hydrogen atom introduced radically new postulates and laid the foundations of quantum mechanics:

- In a hydrogen atom, an electron could revolve in certain stable orbits (called stationary orbits) without the emission of radiant energy.
- The stationary orbits are those for which the angular momentum is some integral multiple of $h/2\pi$. (Bohr's quantisation condition.)
- The third postulate states that an electron might make a transition from one of its specified non-radiating orbits to another of lower energy. When it does so, a single photon is emitted having energy equal to the energy difference between the initial and final states. The frequency of the emitted photon is then given by

$$h\nu = E_i - E_f.$$

Postulate (b) is equivalent to saying that in a stationary state, the circumference of a circular orbit contains integral number of de Broglie wavelengths.

$$2\pi r = n\lambda = n \frac{h}{mv}, \text{ i.e., } L = mvr = n \frac{h}{2\pi}$$

11. For a circular orbit

$$\frac{mv^2}{r} = \frac{e^2}{4\pi\epsilon_0 r^2}$$

The quantisation condition on angular momentum gives

$$v = \frac{nh}{2\pi mr}; n = 1, 2, 3, 4, \dots$$

These equations give

$$r = \left(\frac{n^2}{m}\right) \left(\frac{h}{2\pi}\right)^2 \left(\frac{4\pi\epsilon_0}{e^2}\right)$$

The total energy E is given by

$$E = \frac{1}{2}mv^2 - \frac{e^2}{4\pi\epsilon_0 r}$$

which on substituting for v and r gives

$$E_n = -\frac{1}{2} \frac{mc^2}{n^2} \left(\frac{e^2}{4\pi\epsilon_0 (h/2\pi)c} \right)^2$$

$$E_n = -\frac{13.6}{n^2} \text{ eV}$$

This formula holds good for elliptic orbits as well.

The Rydberg formula for atomic hydrogen spectrum is given by

$$\nu_y = \frac{me^4}{8\epsilon_0^2 h^3} \left(\frac{1}{n_f^2} - \frac{1}{n_i^2} \right)$$

Lyman series (in UV region) corresponds to

$$n_f = 1 ; n_i = 2, 3, 4, \dots$$

Balmer series (in visible region) corresponds to

$$n_f = 2 ; n_i = 3, 4, 5, \dots$$

12. Bohr's model for hydrogen atom fails for atoms having more than one electron. Most simplified version of Bohr's model for multi-electron atoms, the central-field model, suggests that in the ground state of a complex atom, all the electrons should be in the ground state. If this be the case then there would be a gradual change in the physical and chemical properties of elements with increasing number of electrons.

A variety of evidence shows that this is *not* what happens at all. The elements exhibit repetitive properties as function of their position in Periodic Table.

The energy required to remove the most loosely bound electron from a neutral atom is called the *ionisation energy*. A plot of this energy of an element as a function of its position in the periodic table exhibits a repetitive character.

The remarkable similarity in the chemical and physical properties of the elements in each vertical column of the Periodic Table are evidence enough that the atoms are constructed according to systematic rules. In a complex atom all the electrons cannot be in the lowest energy state.

13. The key to this puzzle, discovered by the Swiss physicist Wolfgang Pauli in 1925, is called the *Pauli exclusion principle*. It states that *no two electrons can occupy the same quantum-mechanical state*.

In the central-field model the quantum-mechanical state of an electron is identified not by a single quantum number n as in the Bohr model, but by a set of quantum numbers, all integers, usually called n , l , and m . The first, n , is called the *principal quantum number*, corresponding to n for the hydrogen atom. The energy of the state and its distance from the nucleus increases with n . The quantum number l designates the magnitude of the angular momentum, L , according to the equation,

$$L^2 = l(l+1) \left(\frac{h}{2\pi} \right)^2$$

The value of l can be zero or any positive integer up to and including $(n - 1)$. The quantum number, m designates the component of the angular momentum in axis direction, usually taken to be the z -axis. Specifically,

$$L_z = \frac{mh}{2\pi}$$

The value of m for any electron can be zero or any positive or negative integer up to and including $\pm l$.

The model leads to the shell structure of atom. The principal quantum number n characterises a shell whereas the quantum number l defines a sub-shell.

The electron is visualised not as a point charge but as a small ball of charge spinning on its axis. Associated with this spin is an angular momentum, and experiment shows that the component of this extra angular momentum in the direction of a specified axis (usually taken as the z -axis) is always one of the two values $\frac{1}{2}(h/2\pi)$ or $-\frac{1}{2}(h/2\pi)$. For a given set of values of n , l , m , there are two choices for the orientation of the spin angular momentum, corresponding to values ± 1 of a *fourth* quantum number s . A state is completely defined by a set of four quantum numbers (n , l , m , s).

When we include electron spin, the Pauli's exclusion principle can be restated as: *In an atom, no two electrons can have all four quantum numbers the same.*

14. When a solid target, such as copper or tungsten is bombarded with electrons whose kinetic energies are in the kilo electron volt (keV) range, X-rays are emitted. X-rays are electromagnetic waves having wavelengths ranging from 0.001 to 1 nm.

The wavelength spectrum of X-rays emitted from a target consists of a broad distribution with a few sharp peaks superimposed upon it.

A prominent feature of the continuous part of the spectrum is that the spectrum does not exist below a sharply defined **cut-off wavelength** λ_{\min} . The cut-off wavelength depends only on the accelerating potential and not on the material of the target anode.

The sharp peaks such as K_{α} and K_{β} form the characteristic spectrum and the corresponding X-rays are called characteristic X-rays. The characteristic X-rays constitute a line spectrum and are characteristic of the target element.

15. In 1913, British physicist H.G.J. Moseley measured the frequencies of characteristic X-rays from a number of elements. He noted that if, for a given spectral line, say, K_{α} , the square root of the frequency is plotted against the position of the element in the Periodic Table, a straight line results. The results obtained by Moseley could be expressed as:

$$\sqrt{\nu} \propto Z$$

where ν is the frequency of the spectral line and Z is the atomic number of the target element. This relation is known as Moseley's law. It allowed him to conclude that the property that determines the position of an element in the Periodic Table is not its atomic mass but its atomic number Z .

16. Laser light arises by stimulated emission. That is, radiation of a frequency given by

$$h\nu = E_x - E_0$$

can cause an atom to undergo a transition from an upper energy level (of energy E_x) to a lower energy level with a photon of frequency ν being emitted. The stimulating photon and the emitted photon are identical in every respect and combine to form laser light.

For emission process to predominate, there must normally be a population inversion; that is, there must be more number of atoms in the upper energy level than in the lower one.

Physical Quantity	Symbol	Dimensions	Units	Remarks
Impact parameter	b	[L]	m	Closest distance of approach to the nucleus
Wavelength	λ	[L]	m	
Rydberg constant	R	[L ⁻¹]	m ⁻¹	
Bohr radius	a_0	[L]	m	Radius of the first Bohr orbit in a hydrogen atom
Atomic number	Z			Total positive charge on the nucleus in units of electronic charge

POINTS TO PONDER

1. Both, the Thomson's as well as the Rutherford's models constitute an unstable system. Thomson's model is unstable electrostatically, while Rutherford's model is unstable because of electromagnetic radiation of orbiting electrons.
2. Bohr gave the concept that atoms exist in discrete quantum states of well-defined energy. However, he also propounded the correspondence principle, which stated that for large enough quantum numbers, the predictions of quantum physics merge smoothly with those of classical physics.
3. Besides the problem of stability and spectra for the classical Rutherford atom, there was another basic problem that worried Neils Bohr a great deal. It was that in the classical model, there was no way to get the typical atomic size which was known to be of the order of 10^{-10} m. Why are atoms not, say, 100 times bigger or 100 times smaller than what they actually are? Classical physics had no answer to this question.

The relevant constants in the problem are mass m and charge e of the electron. If you play with using dimensional analysis, you will find that with m , e , and ϵ_0 alone, you cannot get any length dimension. But if you combine m , e , ϵ_0 with Planck's constant h , you can get a length dimension

$$\left(\frac{h^2 \epsilon_0}{m e^2} \right). \text{ Numerically, this is close to the scale of atomic size. Bohr might}$$

have played this dimensional game to guess that h must play a role in his theory.

4. What made Bohr quantise angular momentum (second postulate) and not some other quantity? Note, h has dimensions of angular momentum, and for circular orbits, angular momentum is a very relevant quantity. The second postulate is then so natural!
5. The orbital picture in Bohr's model of the hydrogen atom was inconsistent with the uncertainty principle. It was replaced by modern quantum mechanics in which Bohr's orbits are regions where the electron may be found with large probability.
6. Why does the characteristic X-ray spectrum show such impressive regularities from element to element whereas the optical spectrum in the visible and near-visible region does not? The key to the identity of an element is the charge on its nucleus. The K-electrons, which play an important role in the production of characteristic X-ray spectrum, lie very close to the nucleus and are thus sensitive probes of its charge.
7. A given energy level may not correspond to just one quantum state. For example, a state is characterised by four quantum numbers (n , l , m , and s), but for a pure Coulomb potential (as in hydrogen atom) the energy depends only on n . For a central potential, in general, energy depends on both n and l .
8. In Bohr model, contrary to ordinary classical expectation, the frequency of revolution of an electron in its orbit is not connected to the frequency of spectral line. The latter is the difference between two orbital energies divided by h . For transitions between large quantum numbers (n to $n - 1$, n very large), however, the two coincide as expected by Bohr's correspondence principle.

EXERCISES

- 13.1** What is the distance of closest approach when a 5.0 MeV proton approaches a gold nucleus?
- 13.2** A 12.5 MeV α -particle approaching a gold nucleus is deflected back by 180° . How close does it approach the nucleus?
- 13.3** 2.3 eV separates two energy levels in an atom. What is the frequency of radiation emitted when the atom transits from the upper level to the lower level?
- 13.4** State the basic postulates of Bohr's model of hydrogen atom. Derive an expression for the stationary energy levels of a hydrogen atom.
- 13.5** The ground state energy of hydrogen atom is -13.6 eV. What are the kinetic and potential energies of the electron in this state?
- 13.6** The radius of the innermost electron orbit of a hydrogen atom is 5.3×10^{-11} m. What are the radii of the $n = 2$ and $n = 3$ orbits?
- 13.7** A 10 kg satellite circles earth once every 2 h in an orbit having a radius of 8000 km. Assuming that Bohr's angular momentum postulate applies to satellites just as it does to an electron in the hydrogen atom, find the quantum number of the orbit of the satellite.
- 13.8** In a neon atom the energies of the 3s and 3p states are, respectively, 16.70 eV and 18.70 eV. What wavelength corresponds to 3p - 3s transition in neon atom? In the helium-neon laser why is this line not observed with the same intensity as the 632.8 nm laser line?
- 13.9** A hypothetical atom has energy levels uniformly separated by 1.3 eV. At a temperature 2500 K, what is the ratio of the number of atoms in 15^{th} excited state to the number in 13^{th} excited state?
- 13.10** A population inversion for two energy levels is often described by assigning a negative Kelvin temperature to the system. What negative temperature would describe a system in which population of the upper energy level exceeds that of the lower by 10% and the energy difference between the two levels is 2.2 eV?
- 13.11** A high-powered laser beam ($\lambda = 650$ nm) with a beam diameter of 0.1 m is aimed at the Moon, 3.8×10^8 m distant. The beam spreads only because of diffraction. The angular location of the edge of the central diffraction disc is given by

$$\sin\theta = \frac{1.22\lambda}{d}$$

where d is the diameter of the beam aperture. What is the diameter of the central diffraction disc on the Moon's surface?

- 13.12** Through what minimum potential must an electron in an X-ray tube be accelerated so that it can produce X-rays with wavelength of 0.050 nm?
- 13.13** Here are the K_α wavelengths of a few elements:

Element	Element Number in Periodic Table	λ (pm)
Ti	22	275
V	23	250
Cr	24	229
Fe	26	193
Co	27	179
Cu	29	154

Make a Moseley plot from these data and verify Eq. (13.32).

- 13.14** In an X-ray tube, the tungsten ($Z = 74$) target is bombarded by electrons.
- What is the minimum value of the accelerating potential that will permit the production of the characteristic K_α and K_β lines of tungsten?
 - For this same accelerating potential, what is λ_{\min} ? The K , L , and M energy levels for tungsten have the energies 69.5, 11.3, and 2.30 keV, respectively.

ADDITIONAL EXERCISES

- 13.15** Choose the correct alternative from clues given at end of the each statement:
- The *size of the atom* in Thomson's model is the atomic size in Rutherford's model. (much greater than/no different from/much less than.)
 - In the ground state of electrons are in stable equilibrium, while in electrons always experience a net force. (Thomson's model/Rutherford's model.)
 - A *classical* atom based on is doomed to collapse. (Thomson's model/Rutherford's model.)
 - An atom has a nearly continuous mass distribution in a but has a highly non-uniform mass distribution in (Thomson's model/Rutherford's model.)
 - The positively charged part of the atom possesses most of the mass in (Rutherford's model/both the models.)
- 13.16** Answer the following questions, which help you understand the difference between Thomson's model and Rutherford's model better.
- Is the average angle of deflection of α -particles by a thin gold foil predicted by Thomson's model much less, about the same, or much greater than that predicted by Rutherford's model?
 - Is the probability of backward scattering (i.e., scattering of α -particles at angles greater than 90°) predicted by Thomson's model much less, about the same, or much greater than that predicted by Rutherford's model?
 - Keeping other factors fixed, it is found experimentally that for small thickness t , the number of α -particles scattered at moderate angles is proportional to t . What clue does this linear dependence on t provide?
 - In which model is it completely wrong to ignore multiple scattering for the calculation of average angle of scattering of α -particles by a thin foil?
- 13.17** A projectile of mass m , charge z , initial speed v and impact parameter b is scattered by a heavy nucleus of charge Z . Use angular momentum and energy conservation to obtain a relation connecting the minimum distance s of the projectile from the nucleus to these parameters. Take $b = 0$, and obtain a formula for the distance of closest approach r_0 [Ignore the size of the nucleus].
- 13.18** For scattering by an 'inverse square' field (such as that produced by a charged nucleus in Rutherford's model) the relation between impact parameter b and the scattering angle θ is given by:

$$b = \frac{Z e^2 \cot(\theta/2)}{4\pi \epsilon_0 (mv^2/2)}$$

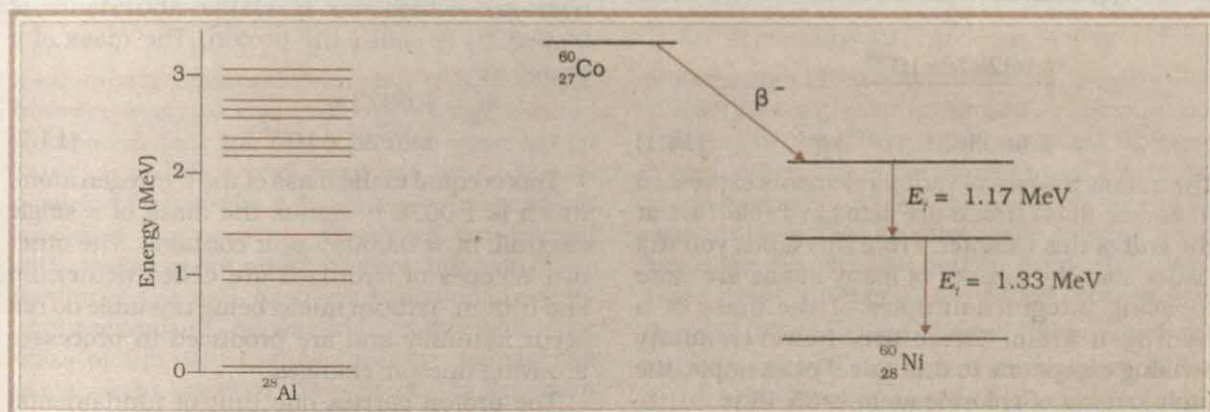
- What is the scattering angle for $b = 0$?
- For a given impact parameter b , does the angle of deflection increase or decrease with increasing energy?
- What is the impact parameter at which the scattering angle is 90° for $Z = 79$ and initial energy of 10 MeV?
- Why is it that the mass of the nucleus does not enter the formula above but the charge does?

- (e) For a given energy of the projectile, does the scattering angle increase or decrease with decrease in impact parameter?
- 13.19** The gravitational attraction between electron and proton in a hydrogen atom is weaker than the coulomb attraction by a factor of about 10^{-40} . An alternative way of looking at this fact is to estimate the radius of the first Bohr orbit of a hydrogen atom if the electron and proton were bound by gravitational attraction. You will find the answer interesting.
- 13.20** Using Bohr's formula for energy quantisation, determine (i) the excitation energy of the $n = 3$ level of He^+ atom. (ii) the ionisation potential of the ground state of Li^{++} atom.
- 13.21** Obtain an expression for the frequency of radiation emitted when a hydrogen atom de-excites from level n to level $(n-1)$. For large n , show that this frequency equals the classical frequency of revolution of the electron in the orbit.
- 13.22** Classically, an electron can be in any orbit around the nucleus of an atom. Then what determines the typical atomic size? Why is an atom not, say, thousand times bigger than its typical size? The question had greatly puzzled Bohr before he arrived at his famous model of the atom that you have learnt in the text. To simulate what he might well have done before his discovery, let us play as follows with the basic constants of nature and see if we can get a quantity with the dimensions of length that is roughly equal to the known size of an atom ($\sim 10^{-10}$ m).
- Construct a quantity with the dimensions of length from the fundamental constants e , m_e , and c . Determine its numerical value.
 - You will find that the length obtained in (a) is many orders of magnitude smaller than the atomic dimensions. Further, it involves c . But energies of atoms are mostly in non-relativistic domain where c is not expected to play any role. This is what may have suggested Bohr to discard c and look for 'something else' to get the right atomic size. Now, the Planck's constant h had already made its appearance elsewhere. Bohr's great insight lay in recognising that h , m_e , and e will yield the right atomic size. Construct a quantity with the dimension of length from h , m_e , and e and confirm that its numerical value has indeed the correct order of magnitude.
- 13.23** A helium atom consists of two electrons orbiting round a nucleus of charge $Z = 2$. But the electrons do not 'see' the full charge $Z = 2$ of the nucleus. Each electron sees the nucleus slightly 'screened' by the other electron so that the effective charge Z_{eff} seen by each electron is less than 2. The ionisation potential for a He atom in its ground state is measured experimentally to be 24.46 V. Estimate the effective charge of the nucleus as seen by each electron in the helium ground state.
- 13.24** A positronium atom is a bound state of an electron (e^-) and its anti-particle, the positron (e^+) revolving round their centre of mass. In which part of the em spectrum does the system radiate when it de-excites from its first excited state to the ground state?
- 13.25** Which state of the triply ionised beryllium (Be^{3+}) has the same orbital radius as that of the ground state of hydrogen? Compare the energies of the two states.
- 13.26** Which level of the doubly ionised lithium (Li^{2+}) has the same energy as the ground state energy of the hydrogen atom? Compare the orbital radii of the two levels.

- 13.27** The total energy of an electron in the first excited state of the hydrogen atom is about -3.4 eV.
- (a) What is the kinetic energy of the electron in this state?
 - (b) What is the potential energy of the electron in this state?
 - (c) Which of the answers above would change if the choice of the zero of potential energy is changed?
- 13.28** Determine the speed of the electron in the $n = 3$ orbit of He^+ . Is the non-relativistic approximation valid?
- 13.29** If Bohr's quantisation postulate (angular momentum $= nh/2\pi$) is a basic law of nature, it should be equally valid for the case of planetary motion also. Why then do we never speak of quantisation of orbits of planets around the sun?
- 13.30** Obtain the first Bohr's radius and the ground state energy of a 'muonic hydrogen atom' (i.e. an atom in which a negatively charged muon (μ^-) of mass about $207 m_e$ orbits around a proton).

CHAPTER FOURTEEN

NUCLEI



14.1 INTRODUCTION

In the previous Chapter we have learnt that in every atom the positive charge is densely concentrated at the center of the atom forming its nucleus. The overall dimensions of a nucleus are much smaller than those of an atom but nevertheless contain most of its total mass. The experiments on scattering of α -particles demonstrated that the radius of a nucleus must be smaller than the radius of an atom by a factor of about 10^4 . In other words, the atom is mostly empty space. The electrons are confined only to a limited part of this empty space.

Just as an atom has a structure, is there a structure to the nucleus too? Is it a continuous mass distribution or does it have further discrete constituents. If so, what is the nature of these constituents and how are they held together? In this Chapter we shall look into some of these aspects and discuss various properties of atomic nuclei.



Marie Skłodowska Curie (1867-1934)

Born in Poland, recognised both as a physicist and as a chemist. The discovery of radioactivity by Henri Becquerel in 1896 inspired Marie and her husband Pierre Curie in their researches and analyses which led to the isolation of radium and polonium elements. She was the first person to be awarded two Nobel Prizes- for Physics in 1903 and for Chemistry in 1911.

14.2 ATOMIC MASSES: COMPOSITION OF NUCLEUS

The mass of an atom is very small; for example the mass of carbon atom, ^{12}C , is 1.992678×10^{-26} kg. It is not very convenient to handle such small quantities, therefore, a mass unit for expressing the atomic masses was introduced. This unit is now defined by taking mass of ^{12}C atom to be 12 atomic mass units (u). According to this definition,

$$\begin{aligned} 1 \text{ u} &= \frac{\text{mass of } ^{12}\text{C atom}}{12} \\ &= \frac{1.992678 \times 10^{-26}}{12} \\ &= 1.660565 \times 10^{-27} \text{ kg} \end{aligned} \quad (14.1)$$

The atomic masses of various elements expressed in atomic mass unit u are listed in Table 14.1 at the end of this Chapter. From this table, you will notice that the masses of many atoms are close to being integral multiples of the mass of a hydrogen atom. There are, however, many striking exceptions to this rule. For example, the atomic mass of chlorine atom is 35.46 u.

A mass spectrometer, which is an improved version of Thomson's apparatus for the measurement of specific charge, e/m , is used for the measurement of atomic masses. The measurement of atomic masses revealed the existence of different varieties of atoms of the same element. The different types of atoms of the same element exhibit similar chemical properties but differ from each other in their masses. Such atomic species of the same element are called *isotopes*. It was discovered that practically every element consists of a mixture of several isotopes. The relative abundance of different isotopes differs from element to element. Chlorine, for example, has two isotopes having masses 34.98 u and 36.98 u, which are nearly integral multiples of the mass of a hydrogen atom. The relative abundances of these isotopes are 75.4 and 24.6 percent, respectively. Thus, the mass of natural chlorine atom is the weighted average of the masses of the two isotopes. That is, mass of natural chlorine atom

$$\begin{aligned} &= \frac{75.4 \times 34.98 + 24.6 \times 36.98}{100} \\ &= 35.47 \text{ u} \end{aligned}$$

which agrees with the atomic mass of chlorine given in Table 14.1. This example thus illustrates that the atomic masses of individual isotopes of an element are close to integers. The observed atomic mass of the element is the weighted average of the atomic masses of the individual isotopes and may not be an integer. It depends on the relative abundance of various isotopes of the element in question.

The lightest element, hydrogen has three isotopes having masses 1.0078 u, 2.0141 u, and 3.0160 u. The nucleus of the lightest atom of hydrogen, which has a relative abundance of 99.985 %, is called the proton. The mass of a proton is

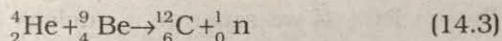
$$\begin{aligned} m_p &= 1.0073 \text{ u} \\ &= 1.6726 \times 10^{-27} \text{ kg}. \end{aligned} \quad (14.2)$$

This is equal to the mass of the hydrogen atom, which is 1.0078 u, minus the mass of a single electron, $m_e = 0.00055$ u, it contains. The other two isotopes of hydrogen are called deuterium and tritium. Tritium nuclei being unstable do not occur naturally and are produced in processes involving nuclear changes.

The proton carries one unit of fundamental charge and is stable. The positive charge in the nucleus is that of the protons. It was earlier thought that the nucleus may contain electrons, but this was ruled out later. Thus, the nucleus does not contain any negative charge. The total number of protons in the nucleus of an atom has, therefore, to be exactly equal to the total number of electrons it contains and hence equal to the atomic number Z .

Since the nuclei of deuterium and tritium are isotopes of hydrogen, they must contain only one proton each. But the masses of the nuclei of hydrogen, deuterium and tritium are in the ratio of 1:2:3. Therefore, the nuclei of deuterium and tritium must contain in addition to a proton, some neutral matter. The amount of neutral matter present in the nuclei of these isotopes, expressed in units of mass of a proton, are approximately equal to one and two, respectively. This fact indicates that the nuclei of atoms, in addition to protons, contain neutral matter in multiples of a basic unit. What is the nature of this neutral matter inside the nucleus? The existence of neutral matter inside the nucleus had been hypothesised by Rutherford in as early as 1920. His arguments, however, were purely

speculative, and until 1932 no experimental evidence supporting them was available. In 1932, Bothe and Becker bombarded beryllium nuclei with α -particles and observed a very penetrating radiation; they then showed that the penetrating radiation was composed of gamma rays. Further study by I. Curie and F. Joliot gave the surprising result (1932) that this 'gamma radiation' also had a component that was capable of imparting energies of several MeV to protons in a cloud chamber. At first, Curie and Joliot interpreted the observed energy transfer as a consequence of scattering of gamma rays from protons. This interpretation, however, was found to be inconsistent with energy and momentum conservation. James Chadwick, however, soon provided the correct explanation. He showed that the recoil protons observed by Curie and Joliot had been hit by a neutral particle of approximately protonic mass, which he called the **neutron**. Chadwick bombarded with neutrons not only hydrogen but also other light nuclei as well. From conservation of energy and momentum, he was able to determine the mass of the new particle '**as very nearly the same as the mass of the proton**'. The emission of a neutron from a beryllium nucleus could be understood as consequence of the reaction



Chadwick was awarded the 1935 Nobel Prize in Physics for his discovery of the neutron.

The mass of a neutron is now known with high degree of accuracy. It is

$$m_n = 1.00866 \text{ u} = 1.6749 \times 10^{-27} \text{ kg} \quad (14.4)$$

A free neutron, unlike a free proton is, unstable. It decays into a proton, an electron and a neutrino (another elementary particle) with a mean life of about 1000 s. It is, however, stable inside the nucleus.

Neutrons and protons are almost identical in the sense that their masses are nearly the same and the force, *nuclear force*, which keeps them together inside the nucleus, does not distinguish them. Therefore, the neutron and the proton have a common name, the *nucleon*. As the proton is positively charged and the neutron is electrically neutral, the electromagnetic force can distinguish them.

The composition of a nucleus can now be described in terms of its constituents, namely, the protons and the neutrons. The number of

protons (called the **atomic number** or **proton number**) is represented by the symbol Z , the number of neutrons (the **neutron number**) by the symbol N . The total number of neutrons and protons in a nucleus is called its **mass number** A , so

$$A = Z + N \quad (14.5)$$

Nuclear species or **nuclides** are represented according to the notation

$${}^A_Z X$$

where X is the chemical symbol of the species. For example, the nucleus of gold is denoted by

${}^{197}_{79}\text{Au}$. It contains 197 nucleons of which 79 are protons and 118 are neutrons.

Nuclides with same atomic number Z but different neutron number N are called **isotopes** of each other. The element gold has 32 isotopes, ranging from $A = 173$ to 204. The isotopes of an element have identical electronic structure and hence the chemical properties. All nuclides with same mass number A are called **isobars**. For example, the nuclides ${}^3_1\text{H}$ and ${}^3_2\text{He}$ are isobars. Nuclides with same neutron number N but different atomic number Z , for example

${}^{198}_{80}\text{Hg}$ and ${}^{197}_{79}\text{Au}$, are called **isotones**.

14.3 SIZE OF THE NUCLEUS

The Geiger-Marsden experiments on the scattering of α -particles from atoms of gold revealed that the size of the gold nucleus has to be less $4 \times 10^{-14} \text{ m}$, which is the distance of closest approach of α -particles of energy 5.5 MeV.

The nucleus, like the atom, is not a solid object like a hard ball. Its 'surface' is not a sharp boundary; still we can assign a size to the nucleus. We can learn about the size and structure of nuclei by bombarding them with a beam of energetic probe particles and observing how the nuclei deflect the incident particles. The particles must be energetic enough so that the associated de Broglie wavelengths are smaller than the nuclear structures they have to probe. Such experiments allow us to assign to each nucleus an effective radius. The effective radius determined depends on the physical quantity being measured. For example, high-energy electron scattering determines the density of charge distribution within the nucleus, whereas scattering experiments involving alpha-particles

determine interaction radius. In the latter case, the nuclear radius obtained is larger than that of the former. However, irrespective of the nature of the probe employed, it is possible to assign to each nucleus an effective radius given

$$r = r_0 A^{1/3} \quad (14.6)$$

in which A is the mass number and r_0 is a constant, which is of the order of the range of nuclear force. The value of the constant r_0 depends on the probe particle, for electrons it is found to be 1.2×10^{-15} m or 1.2 fm. We can see that the volume of the nucleus, which is proportional to r^3 , is directly proportional to A and is independent of the separate values of N and Z . Since neutron and proton are roughly of the same mass, the mass number A is directly proportional to the mass of the nucleus. Thus, the density of nuclear matter is independent of the size of the nucleus just as the density of a liquid is independent of the size of its drop. Nuclear matter in some way behaves like a liquid of constant density. We can compute the density of nuclear matter. For example, for an iron nucleus we have

$$m_{\text{Fe}} = 55.85 \text{ u} = 9.27 \times 10^{-26} \text{ kg}$$

$$A = 56$$

$$\rho_{\text{nuclear}}$$

$$= \frac{9.27 \times 10^{-26}}{(4\pi/3)(1.2 \times 10^{-15})^3} \times \frac{1}{56} \text{ kg m}^{-3}$$

$$= 2.23 \times 10^{17} \text{ kg m}^{-3},$$

which is very large as compared to the density of ordinary matter. The density of matter in neutron stars is comparable to the nuclear density.

14.4 NUCLEAR BINDING ENERGIES

In Section 14.2 we have seen that the nucleus is made up of neutrons and protons. Therefore, it is expected that the mass of the nucleus is equal to the total mass Σm of its individual protons and neutrons. However, the nuclear mass M is found to be always less than Σm . For example, let us take up the case of $^{16}_8\text{O}$; this nucleus has 8 neutrons and 8 protons. We have,

$$\text{Mass of 8 neutrons} = 8 \times 1.00864 \text{ u}$$

$$\text{Mass of 8 protons} = 8 \times 1.00727 \text{ u}$$

$$\text{Mass of 8 electrons} = 8 \times 0.00055 \text{ u}$$

$$\text{Atomic mass of } ^{16}_8\text{O} = 16.0000 \text{ u}$$

$$\begin{aligned} \text{The mass of the } ^{16}_8\text{O nucleus} &= \text{atomic mass} \\ &\text{of } ^{16}_8\text{O} - \text{mass of 8 electrons} \\ &= (16.00000 - 8 \times 0.00055) \text{ u} \\ &= 15.9956 \text{ u} \end{aligned}$$

$$\begin{aligned} \text{Total mass of the constituents of the } ^{16}_8\text{O nucleus} \\ &= (8 \times 1.00864 + 8 \times 1.00727) \text{ u} \\ &= 16.12732 \text{ u} \end{aligned}$$

Thus, we find that the mass of the $^{16}_8\text{O}$ nucleus is less than the total mass of its constituents by 0.13176 u. The difference in mass of a nucleus and its constituents is called the *mass defect*,

$$\Delta M = [Zm_p + (A - Z)m_n] - M \quad (14.7)$$

Using Einstein's mass energy relation,

$$E = mc^2 \quad (14.8)$$

we can express this mass difference in terms of energy as

$$\Delta E_b = \Delta M c^2 \quad (14.9)$$

The Eq. (14.9) shows that if a certain number of neutrons and protons are brought together to form a nucleus of a certain charge and mass, an energy ΔE_b will be released in the process. The energy ΔE_b is called the *binding energy* of the nucleus. If we separate a nucleus into its nucleons, we would have to transfer a total energy equal to ΔE_b , to those particles. Although we cannot tear apart a nucleus in this way, the nuclear binding energy is still a convenient measure of how well a nucleus is held together. A better measure of the binding between the constituents of the nucleus is the **binding energy per nucleon**, ΔE_{bn} , which is the ratio of the binding energy ΔE_b of a nucleus to the number of the nucleons, A in that nucleus

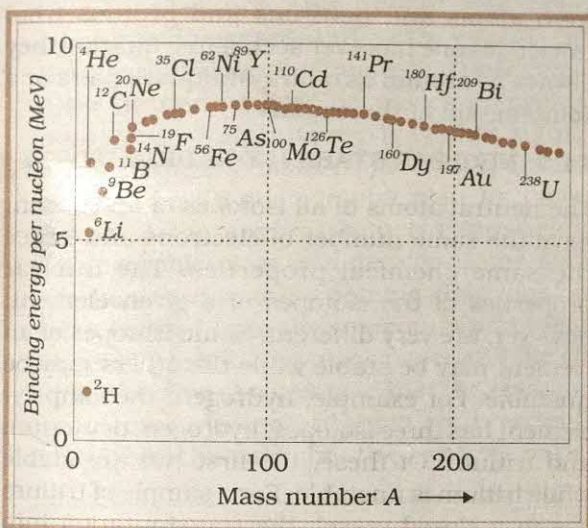
$$\Delta E_{bn} = \Delta E_b / A \quad (14.10)$$

We can think of binding energy per nucleon as the average energy needed to separate a nucleus into its individual nucleons. It is a measure of the strength of the force, which binds the nucleons together inside a nucleus. To get a better insight into the physical significance of this parameter, let us study its variation from nucleus to nucleus.

Figure 14.1 is a plot of the binding energy per nucleon, ΔE_{bn} versus the mass number A for a

large number of nuclei. In this plot we notice the following:

- The binding energy per nucleon, ΔE_{bn} is approximately equal to 8 MeV for nuclei of middle mass number. It is independent of the size of the nucleus.
- ΔE_{bn} is lower for both very light nuclei ($Z \leq 10$) and very heavy nuclei ($Z \geq 70$).



volume. The nuclear force must also be of short range because its influence does not extend far beyond the nuclear 'surface'. It does not depend on charge or is charge independent. It acts between a pair of neutrons, which are electrically neutral. It also acts between a neutron and proton pair and between a pair of protons with equal strength.

The graph of potential energy of a pair of nuclear particles as function of their separation is roughly as shown in Fig. 14.3. This figure depicts the short-range character of nuclear force, of the order of 2 to 3 fm. It is attractive but becomes strongly repulsive when the separation is less than about 1 fm (this region is known as the hard core).

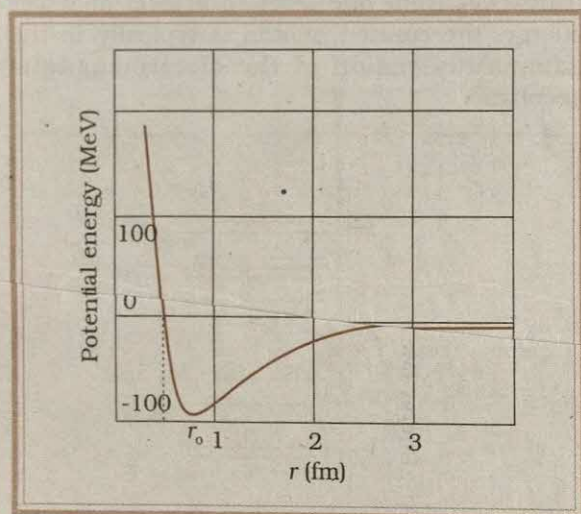


Fig. 14.3 Potential energy of a pair of nucleons as a function of their separation. For a separation greater than r_0 , the force is attractive and for separations less than r_0 , the force is strongly repulsive. The attractive force is strongest when the separation is about 1 fm.

However, the present view is that the nuclear force that binds neutrons and protons in the nucleus is not a fundamental force of nature but is a secondary, or 'spillover' effect of the **strong force** that binds **quarks** together to form neutrons and protons. In much the same way, the attractive force between neutral molecules is a spillover effect of the Coulomb electric force that acts within each molecule to bind it together.

Thus, neutrons and protons are not the ultimate constituents of matter. When protons

and neutrons are subject to collisions with energy of many GeV ($1 \text{ GeV} = 10^9 \text{ eV}$), their behaviour is explained by describing them as made up of quarks. Quarks, like electrons, are thought to be indivisible and among the fundamental building blocks of matter. There are six different kinds of quarks, but they cannot be released in the same way as the electrons from atoms and neutrons and protons from nuclei. No one has ever seen a free quark – they always go around as quark-antiquark pairs or a combination of three quarks.

14.7 NUCLEAR STABILITY-RADIOACTIVITY

The neutral atoms of all isotopes of an element have the same number of electrons and hence the same chemical properties. The nuclear properties of the isotopes of a given element, however, are very different. Some isotopes of an element may be stable while the others may be unstable. For example, hydrogen, the simplest element has three isotopes, hydrogen, deuterium and tritium. Of these, the first two are stable while tritium is unstable. For a sample of tritium gas in a closed vessel, the transmutation into ${}^3_2\text{He}$ occurs smoothly, and the concentration of ${}^3_2\text{He}$ gradually builds up as tritium disappears. After about 12 years, half of a sample of tritium is converted into ${}^3_2\text{He}$. Thus, out of the various isotopes of an element, some may be stable while the others may be unstable.

We organise the nuclides on a **nuclidic chart** like the one shown in Fig. 14.4, in which a nuclide is represented by plotting its proton number against its neutron number. In this plot, stable nuclides are represented by black, and unstable by grey colour. It may be noted that the stable nuclides form a well-defined band and the unstable nuclides lie above and below this band. Note that light stable nuclides tend to lie close to the line $N = Z$, which means that they have about the same numbers of protons and neutrons. Heavier stable nuclides, however, tend to have more neutrons than protons. Note that for a given neutron number, the nuclides below the stable band have less protons and those lying above have more protons as compared to the stable nuclides. Thus, the stability of a nuclide is intimately connected to the relative number of neutrons and protons in that nuclide.

An unstable nuclide spontaneously emits a particle, without the stimulus of any outside agency, transforming itself into a different nuclide. Such a nuclide is said to be *radioactive* and the process of transformation is termed as the *radioactive decay*. *Radioactivity* is the generic name for this process. A.H. Becquerel discovered radioactivity in 1896. While studying the fluorescence and phosphorescence of compounds irradiated with visible light, Becquerel observed an interesting phenomenon. After illuminating some pieces of uranium-potassium sulphate with visible light, he wrapped them in black paper and separated the package from a photographic plate by a piece of silver. After several hours exposure, the photographic plate was developed and showed blackening due to something that must have been emitted from the compound and was able to penetrate both the black paper and the silver.

Rutherford showed later that the emanations given by uranium sulphate were capable of ionising the air in the space between

two oppositely charged metallic plates (an ionisation chamber). The current registered by a galvanometer connected in series with the circuit was taken as the measure of 'activity' of the compound.

A systematic study of the activity of various elements and compounds led Madam Curie to the conclusion that this activity was an atomic phenomenon. She and her husband, Pierre Curie, found that 'ionisability' or 'activity' was associated not only with uranium but also with two other elements that they discovered – radium and polonium. The activity of radium was found to be more than a million times than that of uranium. Since the pioneering work of the Curies, many more radioactive substances have been discovered. Some of them occur naturally while others have been produced artificially in the laboratory.

The activity of radioactive material has been shown to be the result of three different kinds of emanations termed as α , β , and γ radiations or rays. The properties of these radiations are very well known and are summarised below.

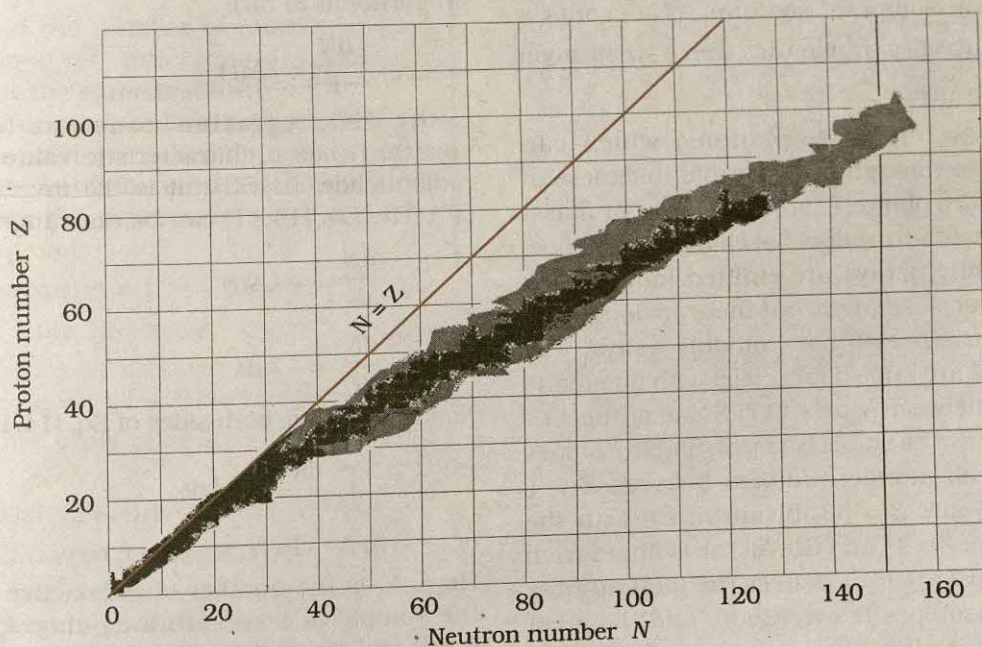


Fig. 14.4 A plot of the known nuclides. The dark shading identifies the band of stable nuclides, the grey shading the radionuclides. Low mass, stable nuclei have essentially equal number of protons and neutrons, but more massive nuclei have increasing number of neutrons. The figure shows that there are no stable nuclei for $Z > 83$.

• α -rays are ${}^4_2\text{He}$ nuclei, emitted from radioactive nuclei are completely stopped by a sheet of paper or by a few centimetres of air. Emission of a ${}^4_2\text{He}$ reduces the mass number of the radionuclide by 4 and its atomic number by 2. For example, the ${}^{208}_{84}\text{Po}$ emits an α -particle, and is transmuted into the stable lead nucleus ${}^{204}_{82}\text{Pb}$.

• β -rays are electrons or particles called positrons that carry a charge $+e$ but are otherwise identical to electrons. The radioactivity is called β^- or β^+ for electrons or positrons, respectively; β^- -rays emitted from radioactive nuclei can penetrate several millimetres of plastic materials. After emission of a β -particle, the mass number of the radioactive nucleus is unchanged but its atomic number is increased or decreased by one. For example, the isotope of potassium ${}^{40}_{19}\text{K}$, emits an electron and is transmuted into a stable calcium isotope ${}^{40}_{20}\text{Ca}$; on the other hand an isotope of sodium, ${}^{22}_{11}\text{Na}$, emits a positron and is transmuted into a stable neon isotope, ${}^{22}_{10}\text{Ne}$.

• γ -rays are energetic photons, which can penetrate through considerable thickness of lead. Since photons carry no charge or mass, emission of a γ -ray does not change the isotope. Frequently, γ -rays are emitted immediately after α - or β -radiation, but there are long-lived radioactive nuclei that emit only γ -rays.

Isotopes are known for nuclei with all atomic numbers between 1 and 117. Some of the 117 elements have no stable isotopes. Stable isotopes occur for all atomic numbers between $Z = 1$ (hydrogen) and $Z = 83$ (bismuth, Bi) with the exception of $Z = 43$ and 61. All the isotopes from $Z = 84$ to 117 are radioactive. The total number of known isotopes is over 2500, and the great majority is radioactive – there are only 266 stable isotopes.

14.7.1 Law of Radioactive Decay

In the previous section we have seen that most of the nuclides that have been identified are

radioactive. A radioactive nuclide spontaneously emits a particle, transforming itself in the process into a different nuclide.

In radioactive decay, as for example in the transmutation of tritium into ${}^3_2\text{He}$, we can predict accurately the amount of tritium remaining at any time, since it follows an exponential law, which we shall derive here. But what happens if we have just one atom of tritium? When is this atom going to turn into ${}^3_2\text{He}$ atom? No one knows: after some time has elapsed, may be it will still be a tritium ${}^3_2\text{He}$ atom, or may be it will have changed into a ${}^3_2\text{He}$ atom. All that can be said is that, after 12.5 years, half the tritium atoms in a large sample will have become ${}^3_2\text{He}$ atoms.

There is absolutely no way to predict whether any given nucleus in a radioactive sample will be among the small number of nuclei that decay during the next second. Each nucleus has the same chance of decay.

Consequently, if a sample contains $N(t)$ radioactive nuclei at any given time t , the rate ($= dN/dt$) at which nuclei will decay is proportional to $N(t)$:

$$-\frac{dN}{dt} = \lambda N(t) \quad (14.11)$$

λ , the **disintegration constant** (or **decay constant**) has a characteristic value for every radionuclide. Its SI unit is the inverse second (s^{-1}). The Eq. (14.11) can be rearranged as:

$$\frac{dN}{dt} = -\lambda N(t)$$

$$\text{or} \quad \frac{dN}{N} = -\lambda dt \quad (14.12)$$

Now, integrating both sides of Eq. (14.12) we get,

$$\int_{N_0}^N \frac{dN}{N} = -\lambda \int_{t_0}^t dt \quad (14.13)$$

$$\text{or} \quad \ln N - \ln N_0 = -\lambda (t - t_0) \quad (14.14)$$

Here, N_0 is the number of radioactive nuclei in the sample at some arbitrary time t_0 . Setting $t_0 = 0$ and rearranging Eq. (14.14) gives us

$$\ln \frac{N}{N_0} = -\lambda t \quad (14.15)$$

which gives

$$N(t) = N_0 e^{-\lambda t} \quad (14.16)$$

in which N_0 is the number of radioactive nuclei in the sample at $t = 0$ and N is the number of radioactive nuclei at any subsequent time t . Note for example, the light bulbs follow no such exponential decay law. If we test 1000 bulbs for their life (time span before they burn out or fuse), we expect that they will 'decay' (that is, burn out) at more or less the same time. The decay of radionuclides follows quite a different law, the **law of radioactive decay** represented by Eq. (14.16).

We are quite often interested in the decay rate R ($= -dN/dt$) than in N itself. Differentiating Eq. (14.16), we find

$$R = -\frac{dN}{dt} = \lambda N_0 e^{-\lambda t}$$

$$\text{Or } R = R_0 e^{-\lambda t} \quad (14.17)$$

an alternative form of the law of radioactive decay [Eq. (14.16)]. Here R_0 is the radioactive decay rate at time $t = 0$, and R is the rate at any subsequent time t . We can now write Eq. (14.11) in terms of the decay rate R of the sample as

$$R = \lambda N \quad (14.18)$$

where R and the number of radioactive nuclei that have not yet undergone decay must be evaluated at the same instant.

The total decay rate R of a sample of one or more radionuclides is called the **activity** of that sample. The SI unit for activity is the **becquerel**, named after the discoverer of radioactivity, Henry Becquerel.

1 becquerel = 1 Bq = 1 decay per second

An older unit, the **curie**, is still in common use:

$$1 \text{ curie} = 1 \text{ Ci} = 3.7 \times 10^{10} \text{ Bq}$$

Some other units of activity in common use are:

$$1 \text{ mCi (milli curie)} = 3.7 \times 10^7 \text{ Bq}$$

$$1 \mu\text{Ci (micro curie)} = 3.7 \times 10^4 \text{ Bq}$$

There are two common time measures of how long any given type of radionuclide lasts. One measure is the **half-life** $T_{1/2}$ of a radionuclide, which is the time at which both N and R have been reduced to one-half their initial values. The other measure is the **mean life** τ , which is the time at which both N and R have been reduced to e^{-1} of their initial values.

To relate $T_{1/2}$ to the disintegration constant λ , we put $R = (1/2)R_0$ in Eq. (14.17) and solving for $T_{1/2}$, we find

$$T_{1/2} = \frac{\ln 2}{\lambda} = \frac{0.693}{\lambda}$$

Similarly, to relate mean life τ to λ , we put $R = e^{-1}R_0$ in Eq. (14.17) and solving for τ , we find

$$\tau = \frac{1}{\lambda}$$

We summarise these results with the following:

$$T_{1/2} = \frac{\ln 2}{\lambda} = \tau \ln 2 \quad (14.19)$$

The half-life or mean life of radioactive nuclei can be as long as the estimated age of the universe, which is 10^{10} years or shorter than 10^{-15} s. Those radioactive elements whose half-life is short are not found in observable quantities in nature today. They have, however, been seen in nuclear reactions. Tritium and plutonium belong to this category.

Example 14.1 The half-life of $^{238}_{92}\text{U}$ against α -decay is 4.5×10^9 years. What is the activity of 1g sample of $^{238}_{92}\text{U}$?

$$\begin{aligned} \text{Answer } T_{1/2} &= 4.5 \times 10^9 \text{ y} \\ &= 4.5 \times 10^9 \text{ y} \times 3.16 \times 10^7 \text{ s/y} \\ &= 1.42 \times 10^{17} \text{ s} \end{aligned}$$

A kmol of an isotope has a mass equal to the atomic mass of that isotope expressed in kg. Hence, 1g of $^{238}_{92}\text{U}$ contains

$$\frac{10^{-3} \text{ kg}}{238 \text{ kg/kmol}} = 4.2 \times 10^{-6} \text{ kmol.}$$

One kmol of any isotope contains Avogadro's number of atoms, and so 1g of $^{238}_{92}\text{U}$ contains $4.20 \times 10^{-6} \text{ kmol} \times 6.025 \times 10^{26} \text{ atoms/kmol} = 25.3 \times 10^{20} \text{ atoms}$. The decay rate R is

$$\begin{aligned} R &= \lambda N \\ &= \frac{0.693}{T_{1/2}} N \end{aligned}$$

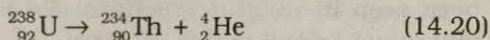
$$\begin{aligned}
 &= \frac{0.693 \times 25.3 \times 10^{20}}{1.42 \times 10^{17}} \text{ s}^{-1} \\
 &= 1.23 \times 10^4 \text{ s}^{-1} \\
 &= 1.23 \times 10^4 \text{ Bq}
 \end{aligned}$$

Example 14.2 Tritium has a half-life of 12.5 y against beta decay. What fraction of a sample of pure tritium will remain undecayed after 25 y.

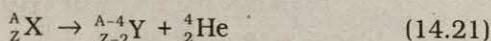
Answer By definition of half-life, $\frac{1}{2}$ of the initial sample will remain undecayed after 12.5 y. In the next 12.5 y, one-half of these nuclei would have decayed. Hence, $\frac{1}{4}$ of the sample of the initial pure tritium will remain undecayed. ◀

14.7.2 Alpha Decay

When a nucleus undergoes **alpha decay**, it transforms to a different nucleus by emitting an alpha particle (a helium nucleus, ${}^4_2\text{He}$). For example, when ${}^{238}_{92}\text{U}$ undergoes alpha decay, it transforms to ${}^{234}_{90}\text{Th}$:



In this process, it is observed that since ${}^4_2\text{He}$ contains two protons and two neutrons, the mass number and the atomic numbers of the daughter nucleus decrease by four and two respectively. Thus, the transformation of a nucleus ${}^A_Z\text{X}$ into a nucleus ${}^{A-4}_{Z-2}\text{Y}$ can be expressed as:



where ${}^A_Z\text{X}$ is the parent nucleus and ${}^{A-4}_{Z-2}\text{Y}$ is the daughter nucleus.

The alpha decay of ${}^{238}_{92}\text{U}$ can occur spontaneously (without an external source of energy) because the total mass of the decay products ${}^{234}_{90}\text{Th}$ and ${}^4_2\text{He}$ is less than the mass of the original ${}^{238}_{92}\text{U}$. Thus, the total mass energy of the decay products is less than the mass energy of the original nuclide. The difference between the initial mass energy and the total mass energy of the decay products is called the Q of the process or the disintegration energy. Thus, the Q of an alpha decay can be expressed as:

$$Q = (m_X - m_Y - m_{\text{He}}) c^2 \quad (14.22)$$

This energy is shared by the daughter nucleus ${}^{A-4}_{Z-2}\text{Y}$, and the alpha particle, ${}^4_2\text{He}$. As the parent nucleus ${}^A_Z\text{X}$ is at rest before it undergoes alpha decay, the alpha-particles are emitted with fixed energy, which can be calculated by applying the principle of conservation of energy and momentum.

Example 14.3 We are given the following atomic masses:

$${}^{238}_{92}\text{U} = 238.05079 \text{ u} \quad {}^4_2\text{He} = 4.00260 \text{ u}$$

$${}^{234}_{90}\text{Th} = 234.04363 \text{ u} \quad {}^1_1\text{H} = 1.00783 \text{ u}$$

$${}^{237}_{91}\text{Pa} = 237.05121 \text{ u}$$

Here the symbol Pa is for the element protactinium ($Z = 91$).

- Calculate the energy released during the alpha decay of ${}^{238}_{92}\text{U}$.
- Show that ${}^{238}_{92}\text{U}$ cannot spontaneously emit a proton.

Answer

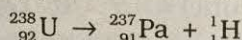
- The alpha decay of ${}^{238}_{92}\text{U}$ is given by Eq. (14.20). The energy released in this process is given by

$$Q = (M_U - M_{\text{Th}} - M_{\text{He}}) c^2$$

Substituting the atomic masses as given in the data, we find

$$\begin{aligned}
 Q &= (238.05079 - 234.04363 - 4.00260) \text{ u} \\
 &\quad \times c^2 \\
 &= (0.00456 \text{ u}) c^2 \\
 &= (0.00456 \text{ u})(931.5 \text{ MeV/u}) \\
 &= 4.25 \text{ MeV.}
 \end{aligned}$$

- If ${}^{238}_{92}\text{U}$ spontaneously emits a proton, the decay process would be



The Q for this process to happen is

$$\begin{aligned}
 &= (M_U - M_{\text{Pa}} - M_{\text{H}}) c^2 \\
 &= (238.05079 - 237.05121 - 1.00783) \text{ u} \\
 &\quad \times c^2 \\
 &= (-0.00825 \text{ u}) c^2 \\
 &= -(0.00825 \text{ u})(931.5 \text{ MeV/u}) \\
 &= -7.68 \text{ MeV}
 \end{aligned}$$

Thus, the Q of the process is negative and therefore it cannot proceed spontaneously. ◀

For a spontaneous alpha decay process Q is always positive [Eq. (14.22)]. However, the half-lives for alpha decay of most of the alpha unstable nuclei are very long. The half-life of $^{238}_{92}\text{U}$ for this decay process is 4.5×10^9 y. Why so long? If $^{238}_{92}\text{U}$ can decay via the process as given in Eq. (14.22), why doesn't every $^{238}_{92}\text{U}$ nuclide in a sample of $^{238}_{92}\text{U}$ atoms simply decay at once? To answer this question, we must examine the process of alpha decay.

We choose a nuclear model in which the alpha-particle is imagined to exist (already formed) inside the nucleus before it escapes from the nucleus. Figure 14.5 shows the approximate potential energy $U(r)$ of the system consisting of the alpha-particle and the residual $^{234}_{90}\text{Th}$ nucleus, as a function of their separation r .

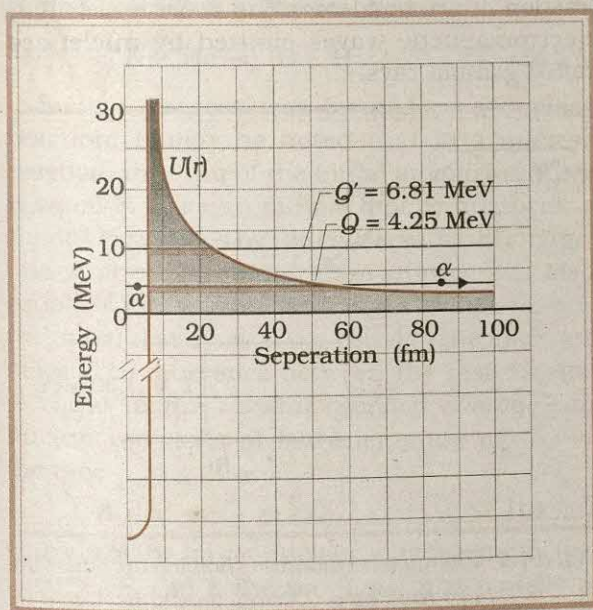


Fig. 14.5 The potential energy function for the emission of an alpha-particle by $^{238}_{92}\text{U}$. The horizontal black line marked at 4.25 MeV shows the disintegration energy for the process. Thick grey portion of this line represents separation r that are classically forbidden to the alpha-particle. The alpha-particle is represented by a dot both inside and outside the potential barrier after the particle has tunneled through.

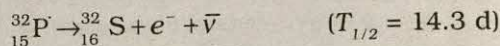
This energy is a combination of (1) the potential energy associated with the (attractive) strong nuclear force that acts in the nuclear interior and (2) a Coulomb potential energy, associated with the (repulsive) electric force that acts between the two particles at all distances.

The horizontal black line marked $Q = 4.25$ MeV shows the disintegration energy for the process. If we assume that this represents the total energy of the alpha-particle during the decay process, then the part of the $U(r)$ curve above this line constitutes a potential energy barrier. This barrier cannot be surmounted classically. If the alpha-particle were at some separation r within the barrier, its potential energy U would exceed its total energy E . This would mean that its kinetic energy would be negative, an impossible situation, classically.

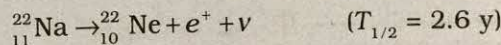
We can see now why the alpha particle is not immediately emitted from the $^{238}_{92}\text{U}$ nucleus. That nucleus is surrounded by an impressive potential barrier, occupying — if you think of it in three dimensions — the volume lying between two spherical shells (of radii about 8 and 60 fm). This argument is so convincing that we now change our question and ask: how can a $^{238}_{92}\text{U}$ nucleus ever emit an alpha-particle? The answer is that there is a finite quantum mechanical probability that a particle can tunnel through an energy barrier that is classically insurmountable. In fact, alpha decay occurs as a result of barrier tunneling.

14.7.3 Beta Decay

A nucleus that decays spontaneously by emitting an electron or a positron is said to undergo **beta decay**. Like alpha decay, this is a spontaneous process, with a definite disintegration energy and half-life. Again like alpha decay, beta decay is a statistical process governed by Eqs. (14.16) and (14.17). In *beta minus* (β^-) decay, an electron is emitted by the nucleus, as in the decay



In *beta plus* (β^+) decay, a positron is emitted by the nucleus, as in the decay of

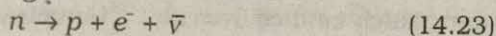


The symbols $\bar{\nu}$ and ν represent antineutrino and neutrino; both are neutral particles, with very little or no mass. These particles are emitted

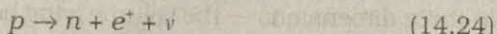
from the nucleus along with the electron or positron during the decay process. Neutrinos interact only very weakly with matter; they can even penetrate the earth without being absorbed. It is for this reason that their detection is extremely difficult and their presence went unnoticed for long.

It may seem surprising that nuclei can emit electrons, positrons and neutrinos, since we have said that nuclei are made of neutrons and protons only. However, we know that atoms emit photons, and we certainly do not say that atoms 'contain' photons. The photons are created during the emission process.

Similarly, the electrons, positrons, and neutrinos are created during the emission process. For beta-minus decay, a neutron transforms into a proton within the nucleus according to



For beta-plus decay, a proton transforms into neutron via



These processes show why the mass number A of a nuclide undergoing beta decay does not change; one of its constituent nucleons simply changes its character according to Eq. (14.23) or (14.24).

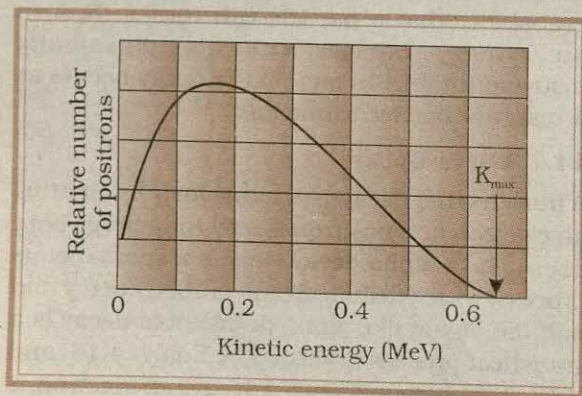


Fig. 14.6 The distribution of the kinetic energies of positrons emitted in the decay of $^{64}_{29}\text{Cu}$.

In both alpha decay and beta decay, the disintegration energy Q is characteristic of the radionuclide. In the alpha decay of a particular radionuclide, every emitted alpha-particle has the same sharply defined kinetic energy. However, in beta decay the disintegration energy, Q , is shared between the three decay products,

the daughter nucleus, electron or positron and the antineutrino or neutrino. As a consequence, the kinetic energy of an electron or a positron in beta decay process is not unique, it may range from zero to a certain maximum K_{\max} as shown in Fig. 14.6. The maximum kinetic energy K_{\max} of an electron or positron must equal the disintegration energy Q . As when the electron or positron carries the maximum energy, the energy carried by the daughter nucleus and the neutrino is approximately zero.

14.7.4 Gamma Decay

In Section 14.5 we introduced the notion of nuclear excited states. Like an excited atom, an excited nucleus can make transitions to a state of lower energy by emitting a photon. As the energy of the nuclear states is of the order of million electron volts (MeV), the photons emitted in transitions between nuclear states can have energy of the order of several MeV. The wavelength of photons of such energy is a fraction of an angstrom. The short wavelength electromagnetic waves emitted by nuclei are called gamma rays.

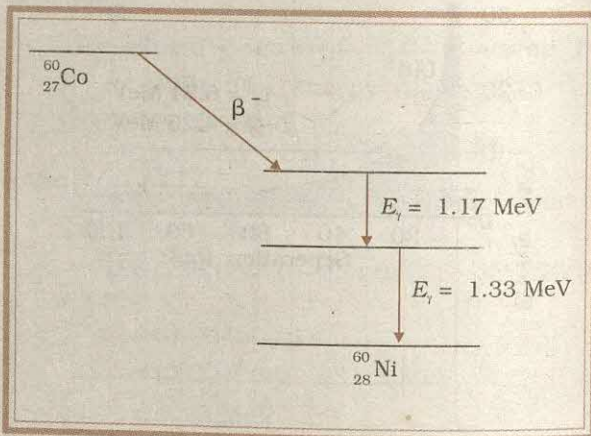


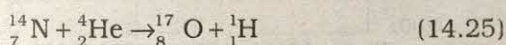
Fig. 14.7 Energy level diagram showing the emission of γ rays by a $^{60}_{27}\text{Co}$ nucleus subsequent to beta decay.

Most radionuclides after an alpha decay or a beta decay leave the daughter nucleus in an excited state. The daughter nucleus by single transition or sometimes by successive transitions reaches the ground state by emitting one or more gamma rays. A well-known example of such a process is that of $^{60}_{27}\text{Co}$. By beta emission, the $^{60}_{27}\text{Co}$ nucleus transforms into

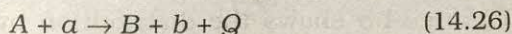
$^{60}_{28}\text{Ni}$ nucleus in its excited state. The excited $^{60}_{28}\text{Ni}$ nucleus so formed then de-excites to its ground state by successive emission of 1.17 MeV and 1.33 MeV gamma rays. This process is depicted in Fig.14.7 through an energy level diagram.

14.8 NUCLEAR REACTIONS

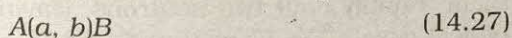
The nuclear disintegrations discussed in Section 14.7 involved natural, spontaneous emission of either a α - or β -particle. Nothing was done to initiate it or nothing could be done to stop it. It occurred to Rutherford in 1919 that it ought to be possible to penetrate a nucleus with a massive high-speed particle such as α -particle and thereby either produce a nucleus with greater mass number or induce an artificial disintegration. Rutherford was successful in bombarding nitrogen with α -particles and obtaining as a result an oxygen nucleus and a proton according to the reaction,



Such a process was termed as a nuclear reaction. It may be noted that in a nuclear reaction, the sum of the initial atomic numbers is equal to the sum of final atomic numbers, a condition imposed by the conservation of charge. The sum of the initial mass numbers is also equal to the sum of final mass numbers, but the initial rest mass is not equal to the final rest mass. The difference between the rest masses is equal to the *nuclear reaction energy*. The nuclear reactions at low energy are mostly of the type



where A is the target nucleus, a is the impinging particle, B and b the products; b is usually a light nucleus or gamma-ray. The reaction represented in Eq. (14.26) is often described in a convenient notation devised by Bothe:



If the reaction energy Q is positive, the reaction is termed as exothermic. On the other hand, if it is negative the reaction is termed as endothermic, and energy is to be provided for in the form of kinetic energy of the impinging particle for the reaction to proceed.

Now, returning to the reaction represented by Eq. (14.25), the rest masses of various particles in u are:

$^4_2\text{He} = 4.00260\text{u}$	$^{17}_8\text{O} = 16.99913\text{u}$
$^{14}_7\text{N} = 14.00307\text{u}$	$^1_1\text{H} = 1.00783\text{u}$
<hr/> 18.00567 u	<hr/> 18.00696 u

The total rest mass of the final products exceeds that of the initial particles by 0.00129 u, which is equivalent to 1.20 MeV. This amount of energy is absorbed in the reaction. If the initial particles did not have this much kinetic energy, the reaction would not have taken place.

Nuclear reactions represented by Eq. (14.26) can be carried out by using a suitable target nucleus and the impinging particle. With the advent of high energy accelerators, α -particle or protons are not the only particles used to investigate nuclear reactions. Accelerated deuteron and light nuclei beams are now available. Such beams are being increasingly employed for various nuclear structure studies. If the nuclear reaction is exothermic, the energy released in the process can be harnessed as a source of energy (*nuclear energy*).

14.8.1 Energy from the Nucleus

When we get energy from coal by burning the fuel in a furnace, we are doing so by tinkering with atoms of carbon and oxygen, rearranging their outer electrons into more stable combinations. When we get energy from uranium in a nuclear reactor, we are again burning a fuel, but then we are tinkering with its nucleus, rearranging its nucleons into more stable configurations.

The electromagnetic Coulomb force holds electrons in atoms; their binding energy is only a few electron volts. However, the nucleons are held in nuclei by the strong force. We have already seen that average binding energy per nucleon in a nucleus is ~ 8 MeV. One kilogram of coal on burning produces $\sim 10^7$ J of energy, on the other hand one kilogram of uranium on fission will generate about 10^{14} J of energy. Thus, for the same quantity of matter, energy released by a nuclear reaction is about a million times that released by a chemical reaction.

In both atomic and nuclear burning, the release of energy is accompanied by a decrease in mass, according to the equation

$$Q = -\Delta mc^2 \quad (14.28)$$

Thus, for the same energy release, a much larger fraction of the available mass (again, by a factor of few million) is consumed in a chemical fuel than in a nuclear fuel.

To have an insight into the process of energy generation from the nucleus, let us have a closer look at the curve of binding energy per nucleon, ΔE_{bn} , shown in Fig. 14.1. It has a long flat region from about $A = 30$ to $A = 170$. In this region, the binding energy per nucleon is nearly constant. However, for $A < 30$ and $A > 170$, the ΔE_{bn} is less than its value in the mid-mass region $30 \leq A \leq 170$. This property of the nuclei in the mid-mass region means that they are more tightly bound in comparison to the nuclei with $A < 30$ and $A > 170$. Therefore, transmutation of less stable nuclei into more tightly bound nuclei through nuclear reactions provides an exciting possibility of releasing nuclear energy.

The nuclear reactions involving nuclei with $A > 170$ and $A < 30$, which can be practical sources of energy are of two broad types (a) fission reactions, and (b) fusion reactions. In the following sections we shall discuss these nuclear reactions.

14.8.2 Nuclear Fission

Soon after the discovery of neutron by Chadwick, Enrico Fermi found that when neutrons bombard various elements, new radioactive elements are produced. He predicted that the neutron, being uncharged, would be a useful nuclear projectile; unlike the proton or alpha-particle, it experiences no repulsive Coulomb force when it nears a nuclear surface. Even *thermal neutrons*, which are slowly moving neutrons in thermal equilibrium with the surrounding matter at room temperature, with a mean kinetic energy of only about 0.04 eV, are useful projectiles in nuclear studies.

In late 1930s, the physicist Lise Meitner and chemists Otto Hann and Fritz Strassmann, working in Berlin and following up the work of Fermi and co-workers, bombarded solutions of uranium salts with such thermal neutrons. They found that after the bombardment a number of new radionuclides were produced. In 1939, after repeated tests, one of the radionuclide was

identified as barium. For Hann and Strassmann, it was rather a mystery as to how this middle-mass element ($Z = 56$) could be produced by the bombardment of uranium ($Z = 92$) with neutrons. Soon after, Meitner and her nephew Otto Frisch solved the puzzle. They suggested the mechanism by which a uranium nucleus, having absorbed a thermal neutron, could split, with the release of energy, into roughly two equal parts, one of which might well be barium. Frisch named this process as **fission**.

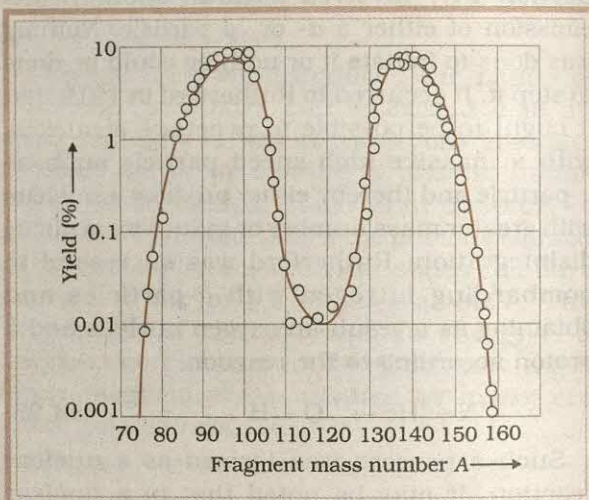
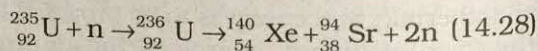


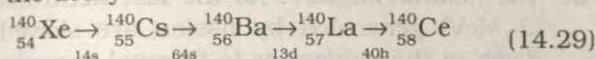
Fig. 14.8 The distribution by mass number of the fragments that are found when many fission events of ^{235}U are examined.

In a typical $^{235}_{92}\text{U}$ fission event, a nucleus $^{235}_{92}\text{U}$ absorbs a thermal neutron, producing a compound nucleus $^{236}_{92}\text{U}$ in a highly excited state. It is this nucleus, which actually undergoes fission, splitting into two fragments. Figure 14.8 shows the distribution by mass number of the fragments produced when $^{235}_{92}\text{U}$ is bombarded with thermal neutrons. The most probable mass numbers occurring in about 7% of the events are centred around $A = 95$ and $A = 140$. The two fragments formed between them, rapidly emit two neutrons, leaving (in a typical case) $^{140}_{54}\text{Xe}$ ($Z = 54$) and $^{94}_{38}\text{Sr}$ ($Z = 38$) as fission fragments. Thus, the overall fission process for this event can be represented by the reaction equation,

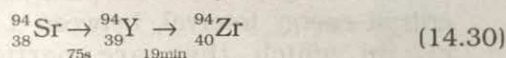


In Eq. (14.28), the fragments $^{140}_{54}\text{Xe}$ and $^{94}_{38}\text{Sr}$ are

both highly unstable, undergoing beta decay until each reaches a stable end product. For Xe the decay chain is



For strontium, the decay chain is



The figure below the arrow indicates the half-life of the decay.

We can estimate the energy released by the fission of a high-mass nuclide by examining the total binding energy per nucleon ΔE_{bn} before and after the fission. Fission can occur because the total mass energy will decrease; that is ΔE_{bn} will increase so that the products of the fission are more tightly bound. Thus, the total energy Q released by the fission is

$$Q = (\text{total final binding energy}) - (\text{initial binding energy}) \quad (14.31)$$

For an estimate, let us assume that fission transforms an initial high-mass nucleon to two middle-mass nuclei with the same number of nucleons. Then we have

$$Q = (\text{final } \Delta E_{bn}) \times (\text{final no. of nucleons}) - (\text{initial } \Delta E_{bn}) \times (\text{initial no. of nucleons}) \quad (14.32)$$

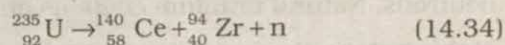
From Fig. 14.1, we see that for high-mass nuclide ($A \approx 240$), the binding energy per nucleon is about 7.6 MeV/nucleon. For the middle-mass nuclides ($A \approx 120$), it is about 8.5 MeV/nucleon. Thus, the energy released by fission of a high-mass nuclide to two middle-mass nuclides is

$$Q = \left(8.5 \frac{\text{MeV}}{\text{nucleon}} \right) (2 \text{ nuclei}) \left(120 \frac{\text{nucleons}}{\text{nucleus}} \right) - \left(7.6 \frac{\text{MeV}}{\text{nucleon}} \right) (240 \text{ nucleons}) \approx 200 \text{ MeV} \quad (14.33)$$

Example 14.4 Find the disintegration energy Q for the fission event represented by Eq. (14.28) taking into account the decay of the fission fragments as displayed in Eqs. (14.29) and (14.30). Some required data are:

$$\begin{aligned} \text{Mass of } {}^{235}_{92}\text{U} &= 235.0439 \text{ u} \\ \text{Mass of n} &= 1.00867 \text{ u} \\ \text{Mass of } {}^{140}_{58}\text{Ce} &= 139.9054 \text{ u} \\ \text{Mass of } {}^{94}_{40}\text{Zr} &= 93.9063 \text{ u} \end{aligned}$$

Answer The key ideas here are (1) that the disintegration energy Q is the energy transferred from mass energy to kinetic energy of the decay products, and (2) that $Q = -\Delta m c^2$, where Δm is the change in mass. Since we have to include the decay products also, therefore, combining Eqs. (14.28), (14.29) and (14.30) we can write the overall transformation as



Only one neutron appears here because one neutron on the left side of the Eq. (14.28) cancels with two neutrons on the right side of the same equation. The mass difference for the reaction in Eq. (14.34) is

$$\begin{aligned} \Delta m &= (139.9054 \text{ u} + 93.9063 \text{ u} + 1.00867 \text{ u}) - (235.0439 \text{ u}) \\ &= -0.22353 \text{ u} \end{aligned}$$

and the corresponding disintegration energy is

$$\begin{aligned} &= -\Delta m c^2 \\ &= (0.22353 \text{ u})(931.5 \text{ MeV/u}) \\ &= 208 \text{ MeV} \end{aligned}$$

which is in good agreement with our estimate in Eq. (14.33). ◀

If the fission event takes place in a bulk solid, most of this disintegration energy, which first goes into the kinetic energy of the decay products, appears eventually as an increase in the internal energy of the system (heat). Five or six percent of the disintegration energy, however, is associated with neutrinos that are emitted during the beta decay of the primary fission fragments. This energy is carried out of the system, as the neutrinos interact very weakly with matter, and is lost.

14.8.3 Nuclear Reactor

In the last section we have seen that fission of a ${}^{235}_{92}\text{U}$ by a thermal neutron leads to release of an extra neutron [Eq. (14.28)]. This extra neutron is then available for initiating fission of another ${}^{235}_{92}\text{U}$ nucleus. In fact, on an average, $2\frac{1}{2}$ neutrons per fission of uranium nucleus are released. The fact that more neutrons are produced in fission than are consumed raises the possibility of a chain reaction with each neutron that is produced triggering another fission. Enrico Fermi first suggested such a possibility in 1939. The chain reaction can be

either uncontrolled and rapid (as in a nuclear bomb) or controlled and steady (as in a nuclear reactor). The first leads to destruction while the latter can be harnessed to generate electric power.

Suppose that we wish to design a reactor based on the fission of ${}^{235}_{92}\text{U}$ by thermal neutrons. Natural uranium contains only 0.7% of this isotope, the remaining 99.3% being ${}^{238}_{92}\text{U}$, which is not fissionable by thermal neutrons. Thus enriched uranium facilitates energy generation by fission. Suppose we have enriched uranium, which contains about 3% ${}^{235}_{92}\text{U}$, still a number of difficulties stand in the way of a working reactor.

1. **Neutron leakage:** Some of the neutrons produced by fission will leak out of the reactor and cease to be part of the chain reaction. Leakage is a surface effect. We can make the fraction of neutrons lost by leakage as small as we wish by making the reactor core large enough, thereby reducing the surface - to - volume ratio.
2. **The neutron energy:** The neutrons produced by fission are fast, with kinetic energies of about 2 MeV. However, fission is induced most effectively by thermal neutrons. Mixing uranium fuel with a substance called moderator can slow the fast neutrons down. A good moderator has two properties. It slows down neutrons by elastic collisions, and it does not remove them from the core by absorbing them. Chadwick's experiments demonstrated that in an elastic collision with hydrogen, the neutron almost comes to rest and proton carries away its energy. Therefore, light nuclei work as most effective moderators for slowing down the neutrons. The moderators commonly used are water, heavy water (D_2O) and graphite. The Apsara reactor at the Bhabha Atomic Research Centre, Mumbai uses water as moderator. The other Indian reactors that are used for power generation use heavy water as moderator. Heavy water is used in reactors using natural uranium as fuel. This is because it has lesser absorption

probability of neutrons than ordinary water.

3. **The neutron capture:** As the fast neutrons (2 MeV) generated by fission are slowed down in the moderator to thermal energies (about 0.04 eV), they must pass through a critical energy interval, between 1 to 100 eV, in which they are particularly susceptible to non-fission capture by ${}^{238}_{92}\text{U}$ nuclei. Such *resonance capture*, which results in the emission of a gamma ray, removes the neutron from fission chain. To minimize such non-fission capture, the uranium fuel and the moderator are not intimately mixed but are 'clumped together', occupying different regions of the reactor volume.

In a typical reactor, the uranium fuel is in the form of uranium oxide pellets, which are inserted end to end into long hollow metal tubes constituting the **fuel rods**. The liquid moderator surrounds the bundle of these rods, forming the reactor **core**. This geometric arrangement increases the probability that a fast neutron, produced in the fuel rod, will find itself in the moderator when it passes through the critical energy interval. Once the neutron has reached thermal energies it may wander back into a fuel rod and produce fission event.

An important reactor parameter is the ratio of number of neutrons present at the beginning of a particular generation to the number present at the beginning of the next generation. This ratio is called the *multiplication factor* (k); it is a measure of the growth rate of the neutrons in the reactor. For $k = 1$, the operation of the reactor is said to be *critical*, which is what we wish it to be for steady power operation. If k becomes greater than one, the reaction rate and the reactor power increases exponentially. Unless the factor k is brought down very close to unity, the reactor will become supercritical and can even explode. In reactor power, inserting control rods into the reactor core controls generation. These rods, containing a material such as cadmium that absorbs neutrons readily, can be inserted farther to reduce the operating power level. They can also be withdrawn to increase the power level or to compensate for the tendency

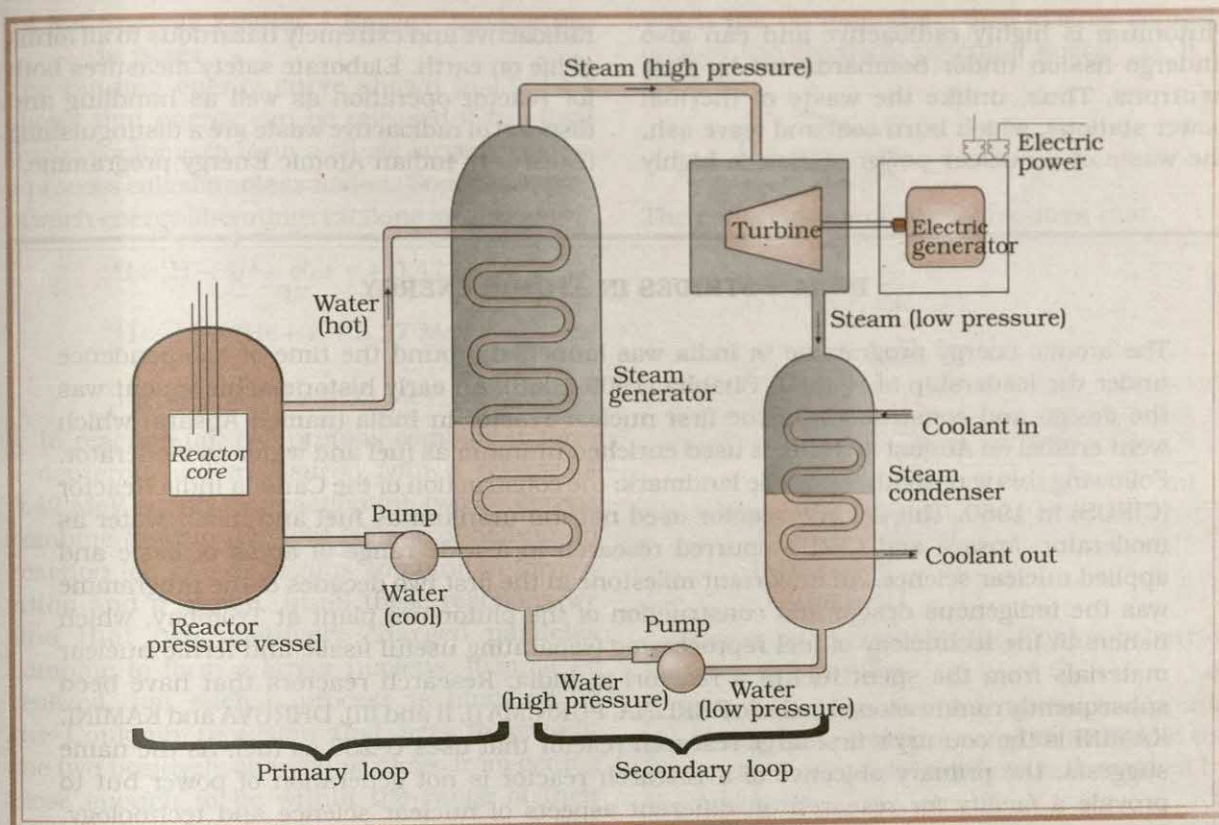


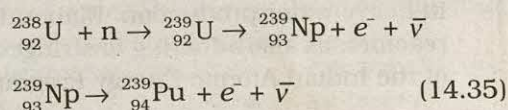
Fig. 14.9 Simplified outlines of a nuclear power plant.

of the reactor to go *subcritical* as (neutron absorbing) fission products build up in the core during continued operation. The control rods thus control the reaction rate.

In addition to control rods, reactors are provided with *safety rods* which when required can be inserted into the reactor and k can be reduced rapidly to less than unity.

The broad outlines of a typical nuclear power plant based on *pressurised-water reactor* are shown in Fig. 14.9. In such a reactor water is used both as the moderator and as the heat transfer medium. In the *primary-loop*, water is circulated through the reactor vessel and transfers energy at high temperature and pressure (at about 600 K and 150 atm) to the steam generator, which is part of the *secondary-loop*. In the steam generator, evaporation provides high-pressure steam to operate the turbine that drives the electric generator. The low-pressure steam from the turbine is cooled and condensed to water and forced back into the steam generator.

As said above, the energy released in nuclear reactions is a million times larger than in chemical reactions. Therefore, the nuclear reactors need fuel a million times less in weight than used in the chemical reactors of the same power capacity. A kilogram of $^{235}_{92}\text{U}$ on complete fission generates about 3×10^4 MW. However, in nuclear reactions highly radioactive elements are continuously produced. Therefore, an unavoidable feature of reactor operation is the accumulation of radioactive waste, including both fission products and heavy *transuranic* elements such as plutonium and americium. The abundant $^{238}_{92}\text{U}$ isotope, which does not fission, on capturing a neutron leads to the formation of plutonium. The series of reactions involved is as follows:



Plutonium is highly radioactive and can also undergo fission under bombardment by slow neutrons. Thus, unlike the waste of thermal power stations, which burn coal and leave ash, the waste of a nuclear power station is highly

radioactive and extremely hazardous to all forms of life on earth. Elaborate safety measures both for reactor operation as well as handling and disposal of radioactive waste are a distinguishing features of Indian Atomic Energy programme.

INDIA'S STRIDES IN ATOMIC ENERGY

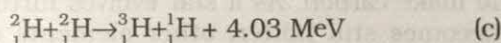
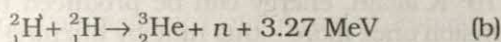
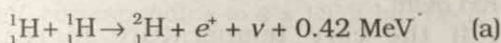
The atomic energy programme in India was launched around the time of independence under the leadership of Homi J. Bhabha (1909-1966). An early historic achievement was the design and construction of the first nuclear reactor in India (named Apsara) which went critical on August 4, 1956. It used enriched uranium as fuel and water as moderator. Following this was another notable landmark: the construction of the Canada India Reactor (CIRUS) in 1960. This 40 MW reactor used natural uranium as fuel and heavy water as moderator. Apsara and CIRUS spurred research in a wide range of areas of basic and applied nuclear science. An important milestone in the first two decades of the programme was the indigenous design and construction of the plutonium plant at Trombay, which ushers in the technology of fuel reprocessing (separating useful fissile and fertile nuclear materials from the spent fuel of a reactor) in India. Research reactors that have been subsequently commissioned include EERLINA, PURNIMA (I, II and III), DHRUVA and KAMINI. KAMINI is the country's first large research reactor that uses U-233 as fuel. As the name suggests, the primary objective of a research reactor is not generation of power but to provide a facility for research on different aspects of nuclear science and technology. Research reactors are also an excellent source for production of a variety of isotopes that find application in diverse fields: industry, medicine and agriculture.

The main objective of the programme from its inception has been to provide safe and reliable electric power for the country's social and economic progress and to be self-reliant in all aspects of nuclear technology. Exploration of atomic minerals in India undertaken since the early fifties has indicated that India has limited reserves in uranium but fairly abundant reserves in thorium. Accordingly, our country has adopted a three-stage strategy of nuclear power generation. The first stage involves the use of natural uranium as a fuel, with heavy water as moderator. The Plutonium-239 obtained from reprocessing of the discharged fuel from the reactors then serves as a fuel for the second stage — the fast breeder reactors. They are so called because they use fast neutrons for sustaining the chain reaction (hence no moderator needed) and, besides generating power, also breed more fissile species (plutonium) than they consume. The third stage, most significant in the long term, involves using fast breeder reactors to produce fissile Uranium-233 from Thorium-232 and to build power reactors based on them.

India is currently well into the second stage of the programme and considerable work has also been done on the third — the thorium utilisation — stage. The country has mastered the complex technologies of mineral exploration and mining, fuel fabrication, heavy water production, reactor design, construction and operation, fuel reprocessing, etc. Pressurised Heavy Water Reactors (PHWRs) built at different sites in the country mark the accomplishment of the first stage of the programme. India is now more than self-sufficient in heavy water production. Elaborate safety measures both in the design and operation of reactors, as also adhering to stringent standards of radiological protection are the hallmark of the Indian Atomic Energy Programme.

14.8.4 Nuclear Fusion - Energy Generation in Stars

The binding energy curve shown in Fig. 14.1 shows that energy can be released if two light nuclei combine to form a single larger nucleus, a process called **nuclear fusion**. Some examples of such energy liberating reactions are as follows:



In reaction (a), two protons combine to form a deuteron and a positron with a release of 0.42 MeV energy. In reaction (b), two deuterons combine to form the light isotope of helium. In reaction (c), two deuterons combine to form a triton and a proton. In all these reactions we find that two positively charged particles combine to form a larger nucleus. It must be realized that such a process is hindered by the Coulomb repulsion that acts to prevent the two positively charged particles from getting close enough to be within the range of their attractive nuclear forces and thus 'fusing'. The height of this *Coulomb barrier* depends on the charges and the radii of the two interacting nuclei. In example 14.5, it is shown that for two protons, the barrier height is 400 keV. The barrier height for more highly charged nuclei is higher.

To generate useful amount of energy, nuclear fusion must occur in bulk matter. What is needed is to raise the temperature of the material until the particles have enough energy - due to their thermal motions alone - to penetrate the Coulomb barrier. This process is called **thermonuclear fusion**.

Example 14.5 Two protons, each having a kinetic energy K , are fired at each other. What must K be if the particles are brought to rest by their mutual Coulomb repulsion? Assume a proton to be a sphere of radius $R = 1 \text{ fm}$.

Answer The initial mechanical energy, E_i of the two protons before collision is given by

$$E_i = 2K$$


When the protons stop, their energy consists only of the electrical potential energy, U . It is given by

$$U = \frac{1}{4\pi\epsilon_0} \frac{e^2}{2R}$$

The conservation of energy requires that

$$2K = \frac{1}{4\pi\epsilon_0} \frac{e^2}{2R}$$

$$\begin{aligned} K &= \frac{e^2}{16\pi\epsilon_0 R} \\ &= \frac{(1.6 \times 10^{-19} \text{ C})^2}{(16\pi)(8.85 \times 10^{-12} \text{ F/m})(1 \times 10^{-15} \text{ m})} \\ &= 5.75 \times 10^{-14} \text{ J} \\ &= 360 \text{ keV} \\ &\approx 400 \text{ keV} \end{aligned}$$

This is approximately the Coulomb barrier between two protons. 

The temperature at which protons in a proton gas would have enough energy to overcome the Coulomb barrier between them is then given by the equation

$$\frac{3}{2} kT = K_{av} \quad (14.36)$$

where K_{av} is the average kinetic energy of the proton, T is the temperature of the proton gas and k is the Boltzmann constant. Equation (14.36) gives

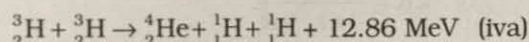
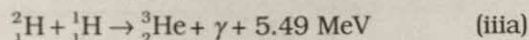
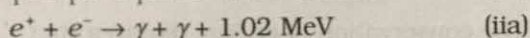
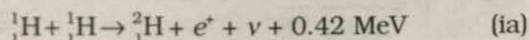
$$\begin{aligned} T &= \frac{2K_{av}}{3k} \\ &= \frac{(2)(5.75 \times 10^{-14} \text{ J})}{(3)(1.38 \times 10^{-23} \text{ J/K})} \\ &= 3 \times 10^9 \text{ K} \end{aligned}$$

The temperature of the core of the Sun is only about $1.5 \times 10^7 \text{ K}$. Therefore, even in the Sun if the fusion is to take place, it must involve protons whose energies are *far* above the average energy.

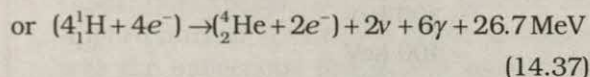
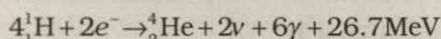
Thus, for thermonuclear fusion to take place, extreme conditions of temperature and pressure are required, which are available only in the interiors of stars. The energy generation in stars takes place via thermonuclear fusion.

The fusion reaction in the Sun is a multi-step process in which hydrogen is burned into

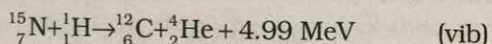
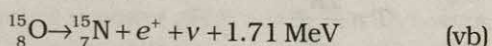
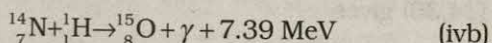
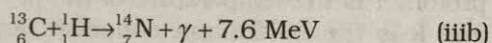
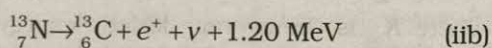
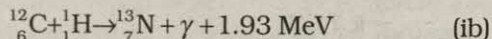
helium, hydrogen being the 'fuel' and helium the 'ashes'. The **proton-proton** (p, p) **cycle** by which this occurs is represented by the following sets of reactions:



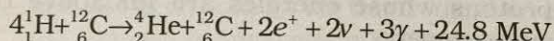
For the fourth reaction to occur, the first three reactions must occur twice, in which case two light helium nuclei unite to form ordinary helium or nucleus. If we consider the combination 2(ia) + 2(iiia) + 2(iiia) + (iva), the net effect is



Thus, four hydrogen atoms combine to form an ${}^4_2\text{He}$ atom with a release of 26.7 MeV of energy. Another set of reactions called the **Carbon cycle** has also been suggested for energy generation in Sun. This set of reaction is represented as follows:



Combining reactions (ib) through (vib) leads to fusion of four hydrogen atoms to form a helium atom, gamma rays, and neutrinos and to liberate about 25 MeV of thermal energy, with the neutrinos escaping. The overall reaction can be written as



It may be noted that in this set of reactions carbon is not destroyed in the process but acts only as a catalyst.

The burning of hydrogen in the Sun's core is alchemy on a grand scale in the sense that one element is turned into another. It has been going on for about 5×10^9 y, and calculations show that there is enough hydrogen to keep the Sun

going for about the same time into the future. In about 5 billion years, however, the Sun's core, which by that time will be largely helium, will begin to cool and the Sun will start to collapse under its own gravity. This will raise the core temperature and cause the outer envelope to expand, turning the Sun into what is called a *red giant*.

If the core temperature increases to 10^8 K again, energy can be produced through fusion once more – this time by burning helium to make carbon. As a star evolves further and becomes still hotter, other elements can be formed by other fusion reactions. However, elements more massive than those near the peak of the binding energy curve of Fig. 14.1 cannot be produced by further fusion.

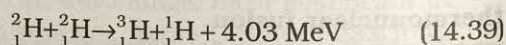
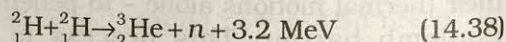
The energy generation in stars takes place via thermonuclear fusion.

14.8.5 Controlled Thermonuclear Fusion

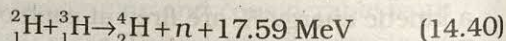
The first thermonuclear reaction on Earth occurred at Eniwetok Atoll on November 1, 1952, when USA exploded a fusion device, generating energy equivalent to 10 million tons of TNT (one ton of TNT on explosion releases 2.6×10^{22} MeV of energy).

A sustained and controllable source of fusion power is considerably more difficult to achieve. It is being pursued vigorously in many countries around the world (including India), because fusion reactor is regarded as the future power source.

The *p-p* cycle discussed in the previous section is not suitable for an Earth-bound fusion reactor as it is extremely slow. The most attractive reaction for terrestrial use appears to be the two deuteron-deuteron reactions,



and the deuteron-triton reaction,



Deuterium, the source of deuterons for these reactions, has an isotopic abundance of only 1 part in about 7000, but is available in unlimited quantity in sea-water.

There are requirements for a successful thermonuclear reactor:

- A High Particle Density:** The (number) density of the interacting particles must be large enough to ensure that the *d-d* collision rate

is high enough. At the high temperatures required, the deuterium would be completely ionised, forming neutral plasma.

- (b) *A High Plasma Temperature:* The plasma must be hot enough to penetrate the Coulomb barrier that tends to keep the interacting particles apart. A plasma ion temperature of 35 keV, corresponding to 4×10^8 K has been achieved in the laboratory.
- (c) *A Long Confinement Time:* A major problem is containing the hot plasma long enough to

maintain it at a density and temperature sufficiently high to ensure the fusion of enough fuel. No solid container can withstand the high temperatures required. Therefore, clever confining techniques, such as magnetic confinement and inertial confinement, are being explored.

Efforts are being made world over to achieve thermonuclear fusion in laboratory. When that happens, humanity will be gifted with a source of unlimited and unpolluted energy.

Table 14.1 Atomic Masses of the Elements

Atomic Number	Element	Symbol	Atomic Mass(u)
1	Hydrogen	H	1.00794
2	Helium	He	4.00260
3	Lithium	Li	6.939
4	Beryllium	Be	9.01218
5	Boron	B	10.811
6	Carbon	C	12.0112
7	Nitrogen	N	14.0067
8	Oxygen	O	15.9994
9	Fluorine	F	18.9984
10	Neon	Ne	20.183
11	Sodium	Na	22.9898
12	Magnesium	Mg	24.312
13	Aluminium	Al	26.9815
14	Silicon	Si	28.0855
15	Phosphorous	P	30.9738
16	Sulphur	S	32.064
17	Chlorine	Cl	35.453
18	Argon	Ar	39.948
19	Potassium	K	39.0983
20	Calcium	Ca	40.08
21	Scandium	Sc	44.955
22	Titanium	Ti	47.88
23	Vanadium	V	50.9415
24	Chromium	Cr	51.996
25	Manganese	Mn	54.9380
26	Iron	Fe	55.847
27	Cobalt	Co	58.933
28	Nickel	Ni	58.71
29	Copper	Cu	63.546
30	Zinc	Zn	65.38

Atomic Number	Element	Symbol	Atomic Mass(u)
31	Gallium	Ga	69.72
32	Germanium	Ge	72.59
33	Arsenic	As	74.921
34	Selenium	Se	78.96
35	Bromine	Br	79.909
36	Krypton	Kr	83.80
37	Rubidium	Rb	85.4678
38	Strontium	Sr	87.62
39	Yttrium	Y	88.9059
40	Zirconium	Zr	91.22
41	Niobium	Nb	92.906
42	Molybdenum	Mo	95.94
43	Technetium	Tc	(99)
44	Ruthenium	Ru	101.107
45	Rhodium	Rh	102.9055
46	Palladium	Pd	106.42
47	Silver	Ag	107.8682
48	Cadmium	Cd	112.41
49	Indium	In	114.82
50	Tin	Sn	118.69
51	Antimony	Sb	121.75
52	Tellurium	Te	127.60
53	Iodine	I	126.9045
54	Xenon	Xe	131.29
55	Caesium	Cs	132.9054
56	Barium	Ba	137.34
57	Lanthanum	La	138.9055
58	Cerium	Ce	140.12
59	Praseodymium	Pr	140.9077
60	Neodymium	Nd	144.24

Atomic Number	Element	Symbol	Atomic Mass(u)
61	Promethium	Pm	(145)
62	Samarium	Sm	150.3
63	Europium	Eu	151.96
64	Gadolinium	Gd	157.25
65	Terbium	Tb	158.9254
66	Dysprosium	Dy	162.50
67	Holmium	Ho	164.9304
68	Erbium	Er	167.26
69	Thulium	Tm	168.9342
70	Ytterbium	Yb	173.04
71	Lutetium	Lu	174.967
72	Hafnium	Hf	178.49
73	Tantalum	Ta	180.9479
74	Tungsten	W	183.85
75	Rhenium	Re	186.207
76	Osmium	Os	190.2
77	Iridium	Ir	192.22
78	Platinum	Pt	195.08
79	Gold	Au	196.9685
80	Mercury	Hg	200.59
81	Thallium	Tl	204.383
82	Lead	Pb	207.2
83	Bismuth	Bi	206.9804
84	Polonium	Po	(209)
85	Astatine	At	(210)

Atomic Number	Element	Symbol	Atomic Mass(u)
86	Radon	Rn	(222)
87	Francium	Fr	(223)
88	Radium	Ra	226.0254
89	Actinium	Ac	227.0278
90	Thorium	Th	232.0381
91	Protactinium	Pa	231.0359
92	Uranium	U	238.0289
93	Neptunium	Np	237.0482
94	Plutonium	Pu	(244)
95	Americium	Am	(243)
96	Curium	Cm	(247)
97	Berkelium	Bk	(247)
98	Californium	Cf	(251)
99	Einsteinium	Es	(254)
100	Fermium	Fm	(257)
101	Mendelevium	Md	(256)
102	Nobelium	No	(259)
103	Lawrencium	Lr	(260)
104	Rutherfordium	Rf	261.11
105	Dubnium	Db	262.114
106	Seaborgium	Sg	263.118
107	Bohrium	Bh	262.12
108	Hassium	Hs	(265)
109	Meitnerium	Mt	(266)

SUMMARY

1. Atoms of every element have a positively charged nucleus. Almost entire mass of the atom is concentrated in the nucleus.
2. Mass of an atom is very small, mass of a carbon atom, ^{12}C , is only 1.992678×10^{-26} kg. To handle such small quantities, a mass unit for expressing the atomic masses was introduced. This unit is now defined by taking mass of ^{12}C atom to be 12 unified atomic mass units (u).

$$1 \text{ u} = 1.660565 \times 10^{-27} \text{ kg.}$$

3. The lightest element, hydrogen has three isotopes having their atomic masses in the ratio of 1:2:3. Therefore, the nuclei of deuterium and tritium must contain in addition to a proton, some neutral matter. The existence of neutral matter inside the nucleus had been hypothesised by Rutherford as early as in 1920. In 1932, Bothe and Becker bombarded beryllium nuclei with α -particles and observed emission of neutral radiation. Using energy momentum conservation, Chadwick concluded in 1932 that this radiation consisted not of photons but neutral particles of approximately protonic mass, which he termed as neutrons.

4. Protons and neutrons are the constituents of a nucleus. The number of protons (called the *atomic number* or *proton number*) is represented by the symbol Z , the number of neutrons (the *neutron number*) by the symbol N . The total number of neutrons and protons in a nucleus is called its *mass number* A .

$$A = Z + N$$

Nuclear species or *nuclides* are represented according to the notation



where X is the chemical symbol of the species. Nuclides with same atomic number Z but different neutron number N are called *isotopes* of each other. All nuclides with same mass number A are *isobars*. Nuclides with same neutron number N but different atomic number Z are called *isotones*.

5. The nucleus, like the atom, is not a solid object with a well-defined surface. Electron-scattering experiments allow us to assign to each nucleus an effective radius given by

$$r = r_0 A^{1/3}$$

in which A is the mass number and r_0 is a constant, which is of the order of the range of nuclear force. The value of the constant r_0 depends on the probe particle, for electrons it is found to be 1.2×10^{-15} m or 1.2 fm. This shows that the density of nuclear matter is independent of A . It is of the order of 10^{17} kg m⁻³.

6. The nuclear mass M is always less than the total mass, Σm , of its constituents. The difference in mass of a nucleus and its constituents is called the *mass defect*,

$$\Delta M = (Z m_p + (A - Z) m_n) - M$$

Einstein's mass energy relation,

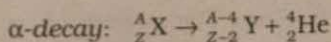
$$E = m c^2$$

can express this mass difference in terms of energy as

$$\Delta E_b = \Delta M c^2$$

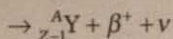
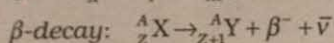
The energy ΔE_b is called the *binding energy* of the nucleus. In the mass number range $A = 30$ to 170, the binding energy per nucleon is nearly constant, about 8 MeV/nucleon.

7. Like an atom, the nucleus exists in nucleonic configurations, which correspond to nuclear stationary states. The stationary state of the lowest energy is called the ground state. The nucleus can be excited from its ground state to stationary states of higher energy. This occurs in processes, which impart energy to the nucleus. When a nucleus makes a transition from one level to a level of lower energy, the emitted photon is typically in the gamma-ray region of the electromagnetic spectrum.
8. Neutrons and protons are bound in a nucleus by the short-range strong nuclear force. The nuclear force does not distinguish between neutron and proton; which are distinguished by electromagnetic force. According to the present view, the nuclear force is not a fundamental force of nature but is a secondary, or 'spillover', effect of the strong force that binds quarks together to form neutrons and protons.
9. The neutral atoms of all isotopes of an element have same chemical properties. However, the nuclear properties of the isotopes of a given element are very different. Some isotopes of an element may be stable while the others may be unstable. The stability of a nuclide is intimately connected to the relative number of neutrons and protons in that nuclide. An unstable nuclide spontaneously emits a particle, without the stimulus of any outside agency, transforming itself into a different nuclide. Such a nuclide is said to be *radioactive* and the process of transformation is termed as the *radioactive decay*. A.H. Becquerel discovered radioactivity in 1896. An unstable nuclide transforms into a stable nuclide by the emission of α -, β -, γ radiations.



The energy released in the process is

$$Q = (M_X - M_Y - M_{\text{He}}) c^2$$



γ -decay: When a nucleus in an excited state makes a transition to state of lower energy, electromagnetic radiation of very short wavelength is emitted. For example, in the β -decay of ${}^{60}_{27}\text{Co}$, ${}^{60}_{28}\text{Ni}$ is formed in an excited state which undergoes transition to the ground state by successive emission of γ -rays of energy 1.17 and 1.33 MeV.

10. Law of radioactive decay:

$$\frac{dN}{dt} = -\lambda N(t)$$

$$\text{or } N(t) = N_0 e^{-\lambda t}$$

where λ is the probability per unit time for a nucleus to decay and $N(t)$ is the number of radioactive nuclei present at time t . Half-life, $T_{1/2}$ is the time in which one-half of the number of nuclei decay.

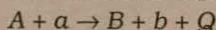
$$T_{1/2} = \frac{\ln 2}{\lambda} = \frac{0.693}{\lambda}$$

The decay rate also called the activity of the sample,

$$R(t) = \lambda N(t)$$

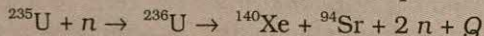
SI unit of activity is *becquerel*, and is equal to one disintegration per second.

11. A nuclear reaction is represented by



where A is the target nucleus, a is the impinging particle, B and b the products. Q is the energy released in the process. First successful nuclear reaction was carried out by Rutherford by bombarding nitrogen nuclei with energetic α -particles resulting in the formation of oxygen nuclei and protons. Nuclear reactions can be employed for the production of newer nuclei or for energy generation.

12. *Nuclear fission*: Splitting of heavy nuclei ($A > 230$) when excited by thermal neutrons into intermediate mass nuclei. For example



where Q is about 200 MeV.

13. *Nuclear reactor*: On an average, 2.5 neutrons are released per fission of uranium. So a chain reaction is possible. An uncontrolled chain reaction leads to explosion as in a nuclear bomb. On the other hand, a controlled chain reaction can be usefully employed for energy generation. A nuclear reactor employs controlled chain reaction to produce energy. Slow neutrons have much greater probability of inducing fission of ${}^{235}_{92}\text{U}$. But the neutrons produced in fission are fast (average energy ~ 2 MeV). The fast neutrons are slowed down by elastic scattering with light nuclei called *moderators* (e.g., water, heavy water and graphite). The reaction rate is controlled by neutron absorbing material (like cadmium rods).

14. *Nuclear fusion*: Two light nuclei combine to form a single larger nucleus with energy liberation. Fusion of hydrogen nuclei into the helium nucleus is the source of energy generation in stars. These reactions require extreme conditions of temperature and pressure so that the reacting nuclei can overcome their electrostatic repulsion. For this reason, these reactions are also termed as *thermonuclear reactions*.

Physical Quantity	Symbol	Dimensions	Units	Remarks
Atomic mass unit		[M]	u	Unit of mass for expressing atomic or nuclear masses. One atomic mass unit equals $1/12^{\text{th}}$ of the mass of ^{12}C atom.
Disintegration or decay constant	λ	$[\text{T}^{-1}]$	s^{-1}	
Half-life	$T_{1/2}$	[T]	s	Time taken for the decay of one-half of the initial number of nuclei present in a radioactive sample.
Activity of a radioactive sample	R	$[\text{T}^{-1}]$	Bq	Measure of the activity of a radioactive source.

POINTS TO PONDER

1. The density of nuclear matter is independent of the size of the nucleus. The mass density of atom does not follow this rule.
2. The binding energy per nucleon, for nuclei of middle mass numbers, is about 8 MeV; a million times higher than the electronic binding energies in an atom.
3. One reason why electrons cannot reside in a nucleus is that the Uncertainty Principle demands that an electron confined to such a small region must have very high energy. The electrostatic attraction between electron and proton, though large at such a small distance, is not enough to bind such a high-energy electron.
4. A free neutron is unstable ($n \rightarrow p + e^- + \bar{\nu}$). But the reverse reaction ($p \rightarrow n + e^+ + \nu$) is not possible since proton is slightly lighter than neutron. Also no other decay is possible for proton because of various conservation laws. So a free proton is stable. Inside a nucleus, both decays are possible, since other nucleons can so share energy and momentum that the conservation of energy and momentum are valid. In a stable nucleus the two processes are in dynamic equilibrium (Exercise 14.20).
5. The nature of nuclear binding energy curve (rising for lighter nuclei, and slightly decreasing for heavier nuclei) shows that exothermic reactions are possible when two light nuclei fuse or when a heavy nucleus undergoes fission into intermediate mass nuclei. For some reason, fusion of light nuclei in stars stops at iron.
6. For fusion, the light nuclei must have sufficient initial energy to cross the Coulomb barrier. Hence, fusion requires high temperatures. However, the actual temperature required is somewhat less than expected classically, because quantum mechanical tunneling of the potential barrier.
7. The Q -value of a nuclear reaction should be obtained from nuclear mass data. However, data are usually in terms of atomic mass. The atomic mass is just the sum of the nuclear mass and the mass of electrons, ignoring the electronic binding energies, which are of the order of magnitude smaller than the energies involved in nuclear reactions.
8. Only in low or medium energy nuclear reactions, the number of protons and number of neutrons are separately conserved. In high energy reactions, protons and neutrons can be converted into other particles. A new quantum number, the Baryon number, is, however, always conserved.

EXERCISES

- 14.1** The three stable isotopes of neon: $^{20}_{10}\text{Ne}$, $^{21}_{10}\text{Ne}$ and $^{22}_{10}\text{Ne}$ have respective abundances of 90.51%, 0.27% and 9.22%. The atomic masses of three isotopes are 19.99 u, 20.99 u and 21.99 u, respectively. Obtain the average atomic mass of neon.

- 14.2** Obtain the binding energy of a nitrogen nucleus ($^{14}_7\text{N}$) from the following data:

$$m_{\text{H}} = 1.00783 \text{ u}$$

$$m_{\text{n}} = 1.00867 \text{ u}$$

$$m_{\text{N}} = 14.00307 \text{ u}$$

Give your answer in MeV.

- 14.3** A given coin has a mass of 3.0 g. Calculate the nuclear energy that would be required to separate all the neutrons and protons from each other. For simplicity assume that the coin is entirely made of $^{63}_{29}\text{Cu}$ atoms (of mass 62.92960 u). The masses of proton and neutron are 1.00783 u and 1.00867 u, respectively.

- 14.4** Obtain the binding energy of the nuclei $^{56}_{26}\text{Fe}$ and $^{209}_{83}\text{Bi}$ in units of MeV from the following data:

$$m_{\text{H}} = 1.007825 \text{ u}$$

$$m_{\text{n}} = 1.008665 \text{ u}$$

$$m(^{56}_{26}\text{Fe}) = 55.934939 \text{ u}$$

$$m(^{209}_{83}\text{Bi}) = 208.980388 \text{ u}$$

Which nucleus has greater binding energy per nucleon?

- 14.5** Write nuclear equations for:

(a) the α -decay of $^{226}_{88}\text{Ra}$

(b) the β^- -decay of $^{32}_{15}\text{P}$

(c) the β^+ -decay of $^{11}_6\text{C}$

- 14.6** A radioactive isotope has a half-life of T years. After how much time is its activity reduced to 6.25% of its original activity?

- 14.7** Obtain the amount of $^{60}_{27}\text{Co}$ necessary to provide a radioactive source of 8.0 mCi strength. The half-life of $^{60}_{27}\text{Co}$ is 5.3 years.

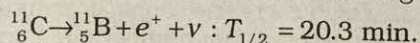
- 14.8** The nucleus of $^{238}_{92}\text{U}$ is unstable against α -decay with a half-life of about 4.5×10^9 years. Write down the equation of the decay and estimate the kinetic energy of the emitted α -particles from the following data:

$$m(^{238}_{92}\text{U}) = 238.05081 \text{ u}$$

$$m(^4_2\text{He}) = 4.00260 \text{ u}$$

$$m(^{234}_{90}\text{Th}) = 234.04363 \text{ u}$$

- 14.9** The radionuclide $^{11}_6\text{C}$ decays according to



The maximum energy of the emitted positron is 0.960 MeV.

Given the mass values:

$$m({}_{6}^{11}\text{C}) = 11.011434 \text{ u}$$

$$m({}_{6}^{11}\text{B}) = 11.009305 \text{ u}$$

$$m_e = 0.000548 \text{ u}$$

calculate Q and compare it with the maximum energy of the positron emitted.

- 14.10** The nucleus ${}_{10}^{23}\text{Ne}$ decays by β^- emission. Write down the β -decay equation and determine the maximum kinetic energy of the electrons emitted. Given that:

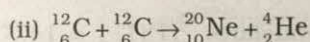
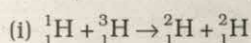
$$m({}_{10}^{23}\text{Ne}) = 22.994466 \text{ u}$$

$$m({}_{11}^{23}\text{Na}) = 22.989770 \text{ u}.$$

- 14.11** The Q value of a nuclear reaction $A + b \rightarrow C + d$ is defined by

$$Q = [m_A + m_b - m_C - m_d]c^2$$

Where the masses refer to nuclear rest masses. Determine from the given data whether the following reactions are exothermic or endothermic.



Atomic masses are given to be

$$m({}_1^1\text{H}) = 1.007825 \text{ u}$$

$$m({}_1^2\text{H}) = 2.014102 \text{ u}$$

$$m({}_1^3\text{H}) = 3.016049 \text{ u}$$

$$m({}_6^{12}\text{C}) = 12.000000 \text{ u}$$

$$m({}_{10}^{20}\text{Ne}) = 19.992439 \text{ u}$$

$$m({}_2^4\text{He}) = 4.002603 \text{ u}.$$

- 14.12** Calculate the disintegration energy Q for the fission of ${}_{42}^{98}\text{Mo}$ into two equal fragments, ${}_{21}^{49}\text{Sc}$. If Q turns out to be positive, explain why this process does not occur spontaneously. Given that:

$$m({}_{42}^{98}\text{Mo}) = 97.90541 \text{ u}$$

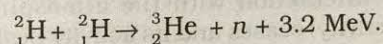
$$m({}_{21}^{49}\text{Sc}) = 48.95002 \text{ u}.$$

$$m_n = 1.00867 \text{ u}$$

- 14.13** The fission properties of ${}_{94}^{239}\text{Pu}$ are very similar to those of ${}_{92}^{235}\text{U}$. The average energy released per fission is 180 MeV. How much energy, in MeV, is released if all the atoms in 1 kg of pure ${}_{94}^{239}\text{Pu}$ undergo fission?

- 14.14** Calculate the height of Coulomb barrier for the head-on collision of two deuterons. The effective radius of deuteron can be taken to be 2.0 fm.

- 14.15** How long an electric lamp of 100 W can be kept glowing by fusion of 2.0 kg of deuterium? The fusion reaction can be taken as

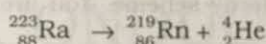
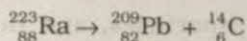


ADDITIONAL EXERCISES

- 14.16** In a Periodic Table the average atomic mass of magnesium is given as 24.312 u. The average value is based on their relative natural abundance on Earth. The three isotopes and their masses are $^{24}_{12}\text{Mg}$ (23.98504 u), $^{25}_{12}\text{Mg}$ (24.98584) and $^{26}_{12}\text{Mg}$ (25.98259 u). The natural abundance of $^{24}_{12}\text{Mg}$ is 78.99% by mass. Calculate the abundances of the other two isotopes.
- 14.17** The neutron separation energy is defined as the energy required to remove a neutron from the nucleus. Obtain the neutron separation energies of the nuclei $^{41}_{20}\text{Ca}$ and $^{27}_{13}\text{Al}$ from the following data:
- $$m_n = 1.008665 \text{ u}$$
- $$m(^{40}_{20}\text{Ca}) = 39.962591 \text{ u}$$
- $$m(^{41}_{20}\text{Ca}) = 40.962278 \text{ u}$$
- $$m(^{26}_{13}\text{Al}) = 25.986895 \text{ u}$$
- $$m(^{27}_{13}\text{Al}) = 26.981541 \text{ u}.$$
- 14.18** A source contains two phosphorous radionuclides $^{32}_{15}\text{P}$ ($T_{1/2} = 14.3 \text{ d}$) and $^{33}_{15}\text{P}$ ($T_{1/2} = 25.3 \text{ d}$). Initially, 10% of the decays come from $^{33}_{15}\text{P}$. How long one must wait until 90% do so?
- 14.19** (a) If the α -decay of $^{238}_{92}\text{U}$ is energetically allowed, what prevents it from decaying all at once? Why is its half-life so long?
 (b) The α -particle faces a Coulomb barrier [see answer to (a) above]. A neutron being uncharged faces no such barrier. Why does the nucleus $^{238}_{92}\text{U}$ not decay spontaneously by emitting a neutron?
- 14.20** (a) The *observed* decay products of a free neutron are a proton and an electron. The emitted electrons are found to have a continuous distribution of kinetic energy with a maximum of $(m_n - m_p - m_e) c^2$. Explain clearly why the presence of a continuous distribution of energy is a pointer to the existence of other unobserved product(s) in the decay.
 (b) If a neutron is unstable with a half-life of about 1000s, why don't all the neutrons of a nucleus decay eventually into protons? How can a nucleus with Z protons and $(A - Z)$ neutrons ever remain stable if the neutrons themselves are unstable?
- 14.21** For the β^+ (positron) emission from a nucleus, there is another competing process known as electron capture (electron from an inner orbit, say, the K -shell, is captured by the nucleus and a neutrino is emitted).
- $$e^- + {}^A_Z\text{X} \rightarrow {}^A_{Z-1}\text{Y} + \nu$$
- Show that if β^+ emission is energetically allowed, electron capture is necessarily allowed but not vice-versa.
- 14.22** The normal activity of living carbon-containing matter is found to be about 15 decays per minute for every gram of carbon. This activity arises from the small proportion of radioactive $^{14}_6\text{C}$ present with the stable carbon isotope $^{12}_6\text{C}$. When the organism is dead, its interaction with the atmosphere (which maintains the above equilibrium activity) ceases and its activity begins to drop.

From the known half-life (5730 years) of $^{14}_6\text{C}$, and the measured activity, the age of the specimen can be approximately estimated. This is the principle of $^{14}_6\text{C}$ dating used in archaeology. Suppose a specimen from Mohenjodaro gives an activity of 9 decays per minute per gram of carbon. Estimate the approximate age of the Indus-Valley civilisation.

- 14.23** Under certain circumstances, a nucleus can decay by emitting a particle more massive than an α -particle. Consider the following decay processes:



- (a) Calculate the Q values for these decays and determine that both are energetically possible.
 (b) The Coulomb barrier height for α -particle emission is 30.0 MeV. What is the barrier height for $^{14}_6\text{C}$? The required data is

$$m(^{223}_{88}\text{Ra}) = 223.01850 \text{ u}$$

$$m(^{209}_{82}\text{Pb}) = 208.98107 \text{ u}$$

$$m(^{219}_{86}\text{Rn}) = 219.00948 \text{ u}$$

$$m(^{14}_6\text{C}) = 14.00324 \text{ u}$$

$$m(^4_2\text{He}) = 4.00260 \text{ u}$$

- 14.24** A 1000 MW fission reactor consumes half of its fuel in 5.00 y. How much $^{235}_{92}\text{U}$ did it contain initially? Assume that all the energy generated arises from the fission of $^{235}_{92}\text{U}$ and that this nuclide is consumed by the fission process.

- 14.25** Consider the fission of $^{238}_{92}\text{U}$ by fast neutrons. In one fission event, no neutrons are emitted and the final stable end products, after the beta-decay of the primary fragments, are $^{140}_{58}\text{Ce}$ and $^{99}_{44}\text{Ru}$. Calculate Q for this fission process. The relevant atomic and particle masses are

$$m(^{238}_{92}\text{U}) = 238.05079 \text{ u}$$

$$m(^{140}_{58}\text{Ce}) = 139.90543 \text{ u}$$

$$m(^{99}_{44}\text{Ru}) = 98.90594 \text{ u}$$

$$m_n = 1.00867 \text{ u}$$

- 14.26** Consider the D-T reaction (deuterium-tritium fusion) given in Eq. (14.40).
 (a) Calculate the energy released in MeV in this reaction from the data:

$$m(^2_1\text{H}) = 2.014102 \text{ u}$$

$$m(^3_1\text{H}) = 3.016049 \text{ u}$$

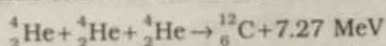
$$m(^4_2\text{He}) = 4.002603 \text{ u}$$

$$m_n = 1.00867 \text{ u}$$

- (b) Consider the radius of both deuterium and tritium to be approximately 1.5 fm. What is the kinetic energy needed to overcome the Coulomb repulsion? To what temperature must the gases be heated to initiate the reaction?

14.27 Calculate and compare the energy released by (a) fusion of 1.0 kg of hydrogen deep within the Sun and (b) the fission of 1.0 kg of ^{235}U in a fission reactor.

14.28 A star converts all its hydrogen to helium, achieving 100% helium composition. It then converts the helium to carbon via the reaction



The mass of the star is $5.0 \times 10^{32} \text{ kg}$, and it generates energy at the rate of $5 \times 10^{30} \text{ W}$. How long will it take to convert all the helium to carbon at this rate?

14.29 Obtain the maximum kinetic energy of β -particles, and the radiation frequencies to γ decays in the following decay scheme. You are given that

$$m({}^{198}\text{Au}) = 197.968233 \text{ u}$$

$$m({}^{198}\text{Hg}) = 197.966760 \text{ u}$$

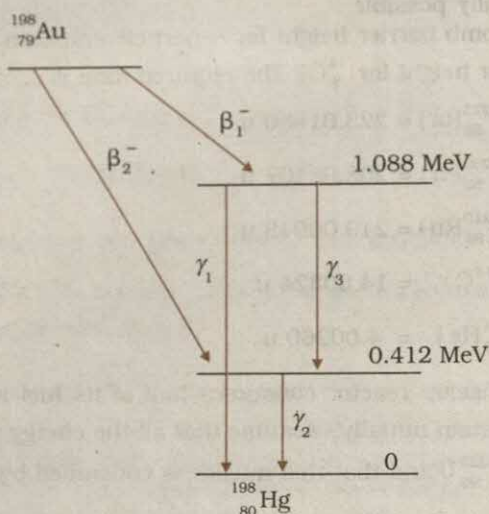


Fig. 14.10

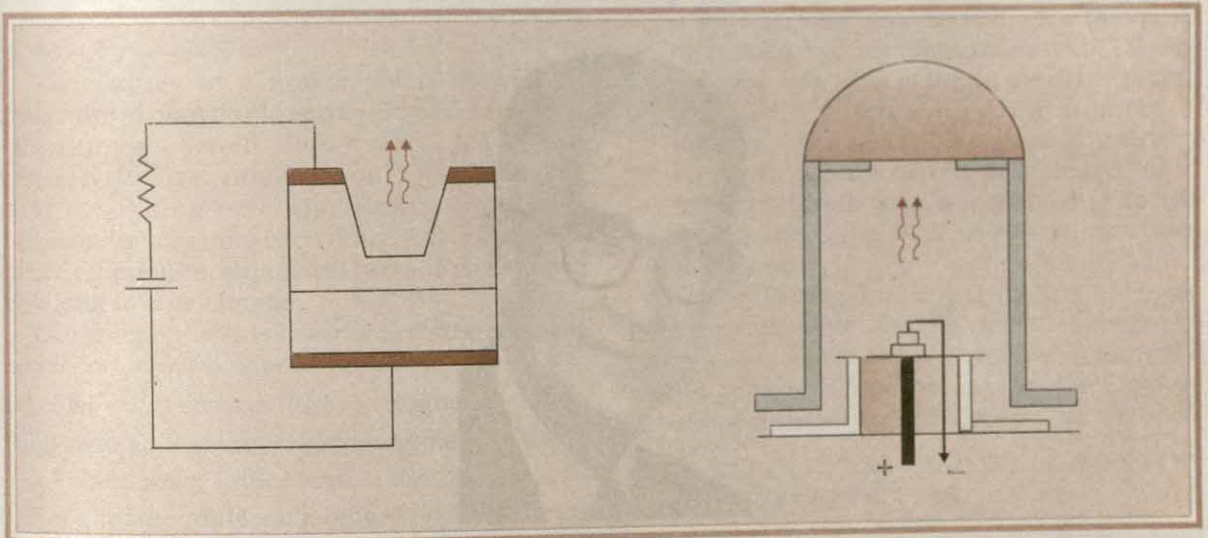
14.30 Answer the following questions:

- Are the equations of nuclear reactions (such as those given in section 14.8) 'balanced' in the sense a chemical equation (e.g., $2\text{H}_2 + \text{O}_2 \rightarrow 2 \text{H}_2\text{O}$) is? If not, in what sense are they balanced on both sides?
- If both the number of protons and the number of neutrons are conserved in each nuclear reaction, in what way is mass converted into energy (or vice versa) in a nuclear reaction?
- A general impression exists that mass-energy interconversion takes place only in nuclear reaction and never in chemical reaction. This is strictly speaking, incorrect. Explain.

14.31 Suppose India had a target of producing by 2000 AD, 100,000 MW of electric power, ten percent of which was to be obtained from nuclear power plant. Suppose we are given that, on an average, the efficiency of utilisation (i.e., conversion to electric energy) of thermal energy produced in a reactor was 25 %. How much amount of fissionable uranium did our country need per year by 2000? Take the heat energy per fission of ^{235}U to be about 200 MeV. Avogadro's number = $6.023 \times 10^{23} \text{ mol}^{-1}$.

CHAPTER FIFTEEN

SEMICONDUCTOR ELECTRONICS: MATERIALS, DEVICES AND SIMPLE CIRCUITS



15.1 INTRODUCTION

Devices in which a controlled flow of electrons can be obtained are the basic **building blocks** of all the electronic circuits. Before the discovery of transistor in 1948, such devices were mostly vacuum tubes or valves like the vacuum diode (consisting of two electrodes, viz. anode/plate and cathode); triode (with three electrodes — cathode, anode and grid); tetrode and pentode (respectively with 4 and 5 electrodes). In a vacuum tube, the electrons are supplied by a heated cathode and the controlled flow of these electrons **in vacuum** is obtained by varying the voltage between its different electrodes. A vacuum is required in the inter-electrode space for the flow of electrons, otherwise the moving electrons may lose their energy on collision with the air molecules in their path. Further, note that the electrons can flow only from the cathode to the anode (i.e., only in one direction). Therefore, such devices are generally referred to as **valves**. These vacuum tube devices are bulky, consume high power, operate generally at high voltages (~100 V), have limited life and low reliability. The seed of the development of modern **solid-state semiconductor electronics** goes back to 1930's as soon as it was realised

that some solid-state semiconductors and their junctions offer the possibility of controlling the number and the direction of flow of charge carriers through them. It is well known that simple excitations like light, heat or small applied voltage can change the number of mobile charges in a semiconductor. Note that the supply and flow of charge carriers in the semiconductor devices are **within the solid itself**, while in the earlier vacuum tubes/valves, the mobile electrons were obtained from a heated cathode and they were made to flow in an **evacuated** space or vacuum. No external heating or large evacuated space is required by the semiconductor devices. These are small in size, consume low power, operate at low voltages, have long life and high reliability. Much before the full implications of the semiconductor devices was formally understood, a naturally occurring crystal *galena* (Lead sulphide, PbS) with a metal point contact attached to it was used as *detector* of radiowaves.

In the following sections, we will introduce the basic concepts of semiconductors and some semiconductor junction diodes (a 2-electrode device) and transistor (a 3-electrode device). A few circuits illustrating their applications are also described.



William Bradford Shockley (1910-1989)

Born in London, England. Educated in California and obtained Ph.D. in 1936 from the Massachusetts Institute of Technology, USA. Shockley's research had been centred in many areas of semi-conductor physics, e.g., energy bands in semi-conductors; order and disorder in alloys; theory of vacuum tubes; theories of dislocations; experiment and theory on ferromagnetic domains; and various topics in transistor physics.

He was awarded the Nobel Prize in Physics in 1956 for his researches on semiconductors and discovery of transistor effect; alongwith John Bardeen (1908 - 1991) and Walter Houser Brattain (1902-1987). This work revolutionised the world of electronics.

15.2 SEMICONDUCTOR PHYSICS: SOME BASICS

On the basis of the relative values of electrical conductivity (σ) or resistivity ($\rho = 1/\sigma$), the solids are broadly classified as:

- (a) **Metals:** They possess very low resistivity (or high conductivity).
 $\rho \sim 10^{-2} - 10^{-8} \Omega \text{ m}$
 $\sigma \sim 10^2 - 10^8 \text{ S m}^{-1}$
- (b) **Insulators:** They have high resistivity (or low conductivity).
 $\rho \sim 10^8 \Omega \text{ m}$
 $\sigma \sim 10^{-8} \text{ S m}^{-1}$
- (c) **Semiconductors:** They have resistivity or conductivity intermediate to metals and insulators.
 $\rho \sim 10^5 - 10^0 \Omega \text{ m}$
 $\sigma \sim 10^{-5} - 10^0 \text{ S m}^{-1}$

The values of ρ and σ given above are indicative of magnitude and could well go outside the ranges as well. Relative values of the resistivity are not the only criteria for distinguishing metals, insulators and semiconductors from each other. There are some other differences, which will become clear as we go along in this chapter.

Our interest in this chapter is in the study of semiconductors which could be:

- (a) *Elemental semiconductors:* Si and Ge
- (b) *Compound semiconductors:* Examples are:
 - Inorganic: CdS, GaAs, CdSe, InP etc.
 - Organic: anthracene, doped phthalocyanines etc.
 - Organic polymers: polypyrrole, polyaniline, polythiophene etc.

Most of the currently available semiconductor devices are based on elemental semiconductors Si or Ge and compound **inorganic** semiconductors. However, after 1990, a few semiconductor devices using organic semiconductors and semiconducting polymers have been developed signalling the birth of a futuristic technology of polymer-electronics and molecular-electronics. In this chapter, we will restrict ourselves to the study of inorganic semiconductors, particularly elemental semiconductors Si and Ge. The general concepts introduced here for discussing the elemental semiconductors by-and-large apply to most of the compound semiconductors as well.

Apart from the above classification of semiconductors on the basis of chemical composition, there is another classification scheme on the basis of the **source** and **nature** of the charge carriers. Such a scheme divides semiconductors as *intrinsic* and *extrinsic* semiconductors discussed below:

- (a) **Intrinsic semiconductor:** These are **pure** semiconducting materials (impurity less than 1 part in 10^{10}). The presence of the mobile charge carriers is the **intrinsic** property of the material. The electrical conduction is by means of mobile electrons and holes (the concept of **hole** as positive charge carriers is introduced later in this chapter).
- (b) **Extrinsic semiconductor:** These are obtained by adding or **doping** the pure semiconductor material with small amounts of certain specific impurities with valency different from that of the atoms of the parent material. Consequently, the number of mobile electrons/holes gets drastically changed. So, the electrical conductivity in such materials is essentially due to the foreign atoms, or in other words, **extrinsic** in nature.

In the following Sections (15.3 and 15.4), we give a qualitative discussion to explain the electrical behaviour of semiconductors. We would mostly concentrate our discussions around the elemental semiconductors Ge or Si.

15.3 ELECTRICAL CONDUCTION IN SEMICONDUCTORS

As you have learnt in Chapter 13, the energy of an electron in an **isolated atom** is decided by the orbit in which it is revolving. In a solid, these electron energies would be different because of the presence of many atoms placed close to each other. This would make the nature of electron motion in a solid different from that in an isolated atom.

There are two ways of describing the phenomena of electrical conduction in semiconductors: (i) Valence-bond model, and (ii) Energy-band model. The valence-bond model is a qualitative approach where we consider the manner in which the valence electrons (viz. the electrons in the outermost orbit) bond themselves to form a solid. This is as we do in **chemistry**. This model gives a **vivid** and **visual picture** of electrical conduction in

semiconductors. The Energy-band model is a more accurate quantum mechanical description.

First, let us consider what happens when atoms bind themselves to form a solid (**Valence bond model**). We know that an isolated atom essentially consists of a nucleus (with net positive charge) and electrons revolving round it in different orbits. The net positive charge in the nucleus is equal to the sum of negative

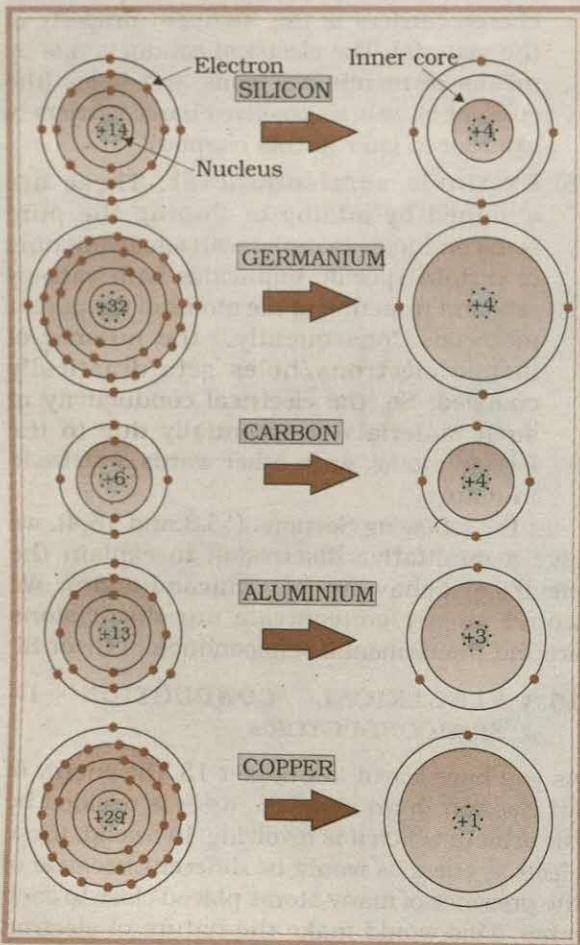


Fig. 15.1 Atomic structure, inner core charges and valence electrons shown schematically for a few elements.

charges on all the orbiting electrons to make the atom electrically neutral. Different orbits can contain only a limited number of electrons. For example, the maximum possible number of electrons in the first, second and third orbits of an atom, respectively, are 2, 8 and 18. As an example, the left hand side of Fig. 15.1 gives the atomic structure of a few elements like

carbon ($Z = 6$); silicon ($Z = 14$); germanium ($Z = 32$); copper ($Z = 29$); aluminium ($Z = 13$). Consider the case of carbon. The first orbit is completely filled with 2 electrons while the second orbit is only partially filled with 4 electrons (it can take a maximum of 8 electrons). These 4 electrons are called **valence electrons** and are responsible for bonding with other atoms. The two inner electrons in the fully filled first orbit (-2 electronic charge) and the carbon nucleus (with $+6$ electronic charge) constitute the *tightly bound inner core* with net $+4$ charge. Hence, we can represent carbon as shown in the right hand side of Fig. 15.1. Similarly, we can see that for silicon, the first and the second orbit are fully filled with 2 and 8 electrons, respectively, and the outermost orbit again has 4 valence electrons. As for carbon, the silicon inner core (nucleus plus inner filled orbits) has again a net $+4$ charge [$(+14) + (-2) + (-8) = +4$] charge. In case of Ge, the inner core consists of $+32$ charges in the nucleus and 2, 8, 18 negative electronic charges in the fully filled first, second and third orbits, respectively. The net inner core charge is $+4$ [$(+32) + (-2) + (-8) + (-18) = +4$] and here also there are 4 valence electrons. Thus, bonding situations for C, Si and Ge are almost similar. Similarly, you can verify for copper and aluminium that the respective number of valence electrons are 1 and 3. Note that some of the above discussed cases like Al or Cu are metals with low resistivity while carbon has high resistivity and Si and Ge behave like semiconductors. Why? You can see that the answer obviously goes beyond the actual number of valence electrons. The C, Si and Ge, all have the same number of valence electrons (4), but their resistivities widely differ. Further, Cu and Al are low resistivity metals even though these have lesser valence electrons (1 and 3) as compared to C, Si or Ge. The answer lies in the manner in which the respective valence electrons of these atoms bond themselves together to form the **solid**. This is qualitatively discussed below.

In a solid, the atoms are held together closely in a well defined 3-dimensional array or lattice by strong forces as the separation between them is quite small (~ 2 to 3 \AA ; nearly twice the radii of the outermost electronic orbits). At such separations, the outer electronic orbits of the neighbouring atoms overlap considerably and hence get significantly distorted. In fact, some

valence electrons may be shared by several atoms and it becomes difficult to say which electron goes with which atom. In conducting metals like Al, Cu etc., the outer orbit electrons are **shared** by all the atoms and the metallic crystal can be visualised as **positive inner cores embedded in a regular fashion in a sea of shared electrons** (Fig. 15.2). Such electrons are

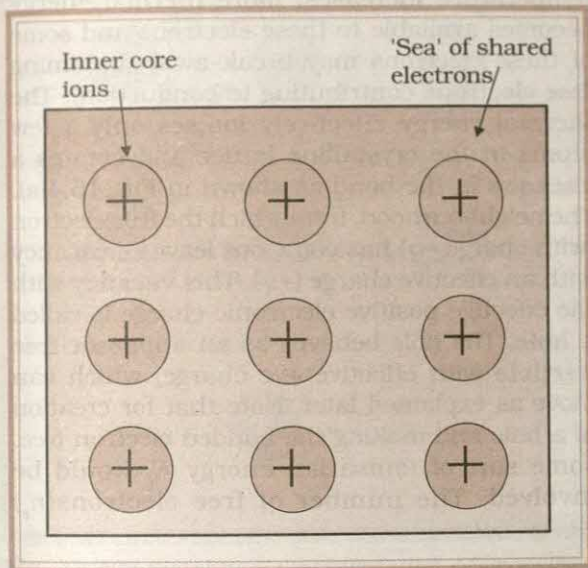


Fig. 15.2 Schematic representation of a metallic crystal showing **sea of shared electrons** in which inner core ions are embedded at fixed positions.

obviously free to travel (hence, termed as **free electrons**) throughout the material randomly. A small electric field results in the flow of these free electrons in the direction of +ve potential and gives low resistivity (or high conductivity) to metals. If we assume that each atom contributes only one electron to the **sea of shared electrons free to move**, even then the number of mobile electrons will be of the order of $\sim 10^{29} \text{ m}^{-3}$ which is very large.

For insulators, the bonding situation is drastically different. The outermost electrons remain **bound** to their parent atoms almost at all temperatures and very high energies are required

for these electrons to breakaway. The number of free electrons is generally less than 10^{14} m^{-3} . Because many mobile charges are not available, the conductivity is low (or resistivity is high). On the contrary, the bonding in semiconductors is such that under thermal excitation (at moderate temperatures) the bonded electrons let themselves loose and can freely wander to give moderate values of conductivity. However, at low temperatures the conductivity of most of the semiconductors is low and comparable to those of insulators. Thus, the semiconductors are materials intermediate between insulators and metals. Since the semiconductor is the material used in most of the modern solid state electronic devices, a detailed description of the **Valence-Bond** situation for a few important semiconductors (particularly Si and Ge) is discussed in the subsequent sub-section.

15.3.1 Valence-Bond Description of Intrinsic Semiconductors

As mentioned earlier, the term **intrinsic semiconductor** refers to **pure** materials in which the conductivity is due to its intrinsic properties rather than the presence of any impurity and/or foreign atom. We shall take the most common case of Ge and Si whose lattice structure is shown in Fig. 15.3(a). These

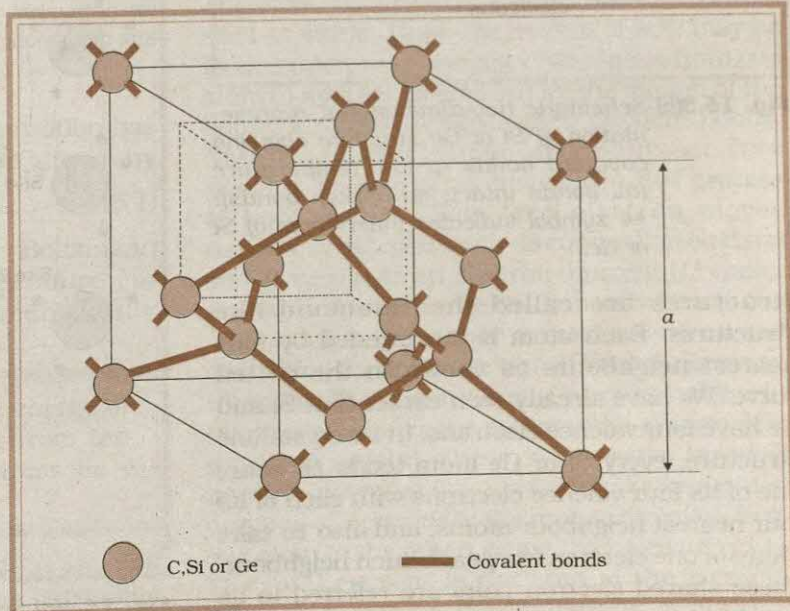


Fig. 15.3(a) Three-dimensional diamond like crystal structure for Carbon, Silicon or Germanium with respective lattice spacing 'a' equal to 3.56, 5.43 and 5.66 Å.

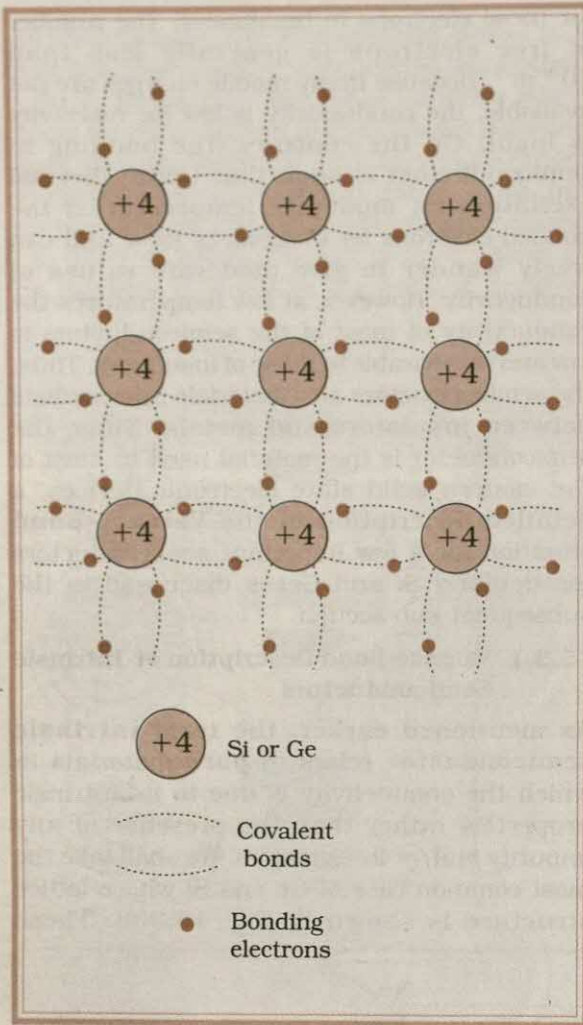


Fig. 15.3(b) Schematic two-dimensional representation of Si or Ge structure showing covalent bonds at low temperature (all bonds intact; no broken bonds). +4 symbol indicates inner cores of Si or Ge.

structures are called the diamond-like structures. Each atom is surrounded by four nearest neighbours as shown in the dotted curve. We have already seen earlier that Si and Ge have four valence electrons. In its crystalline structure, every Si or Ge atom tends to **share** one of its four valence electrons with each of its four nearest neighbour atoms, and also to **take share** of one electron from each such neighbour. These shared electron pairs are referred to as forming a **covalent bond** or simply a **valence bond**. The two shared electrons can be assumed to shuttle back-and-forth between the associated

atoms holding them together strongly. Fig. 15.3(b) schematically shows the 2-dimensional representation of Si or Ge structure shown in Fig. 15.3(a) which overemphasises the covalent bond. The Fig. 15.3(b) is an idealised picture in which no bonds are broken (all bonds are intact). Such a situation arises at low temperature. As the temperature increases, more thermal energy becomes available to these electrons and some of these electrons may break-away (becoming **free** electrons contributing to conduction). The thermal energy effectively ionises only a few atoms in the crystalline lattice and creates a **vacancy** in the bond as shown in Fig. 15.4(a). The neighbourhood, from which the free electron (with charge $-q$) has come out leaves a vacancy with an effective charge $(+q)$. This **vacancy** with the effective positive electronic charge is called a **hole**. The hole behaves as an **apparent free particle** with effective +ve charge, which can move as explained later. Note that for creation of a hole and making the bonded electron free, some sort of ionisation energy E_g would be involved. The number of free electrons (n_e)

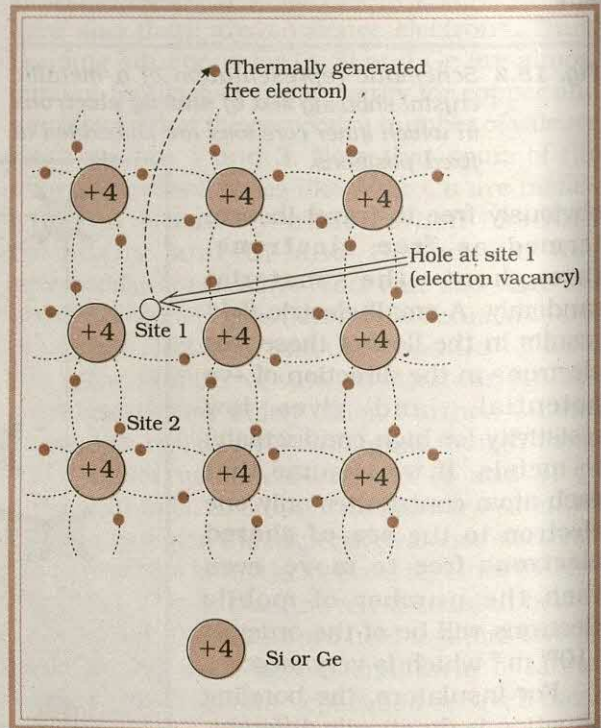


Fig. 15.4(a) Schematic model of generation of hole 'h' at site 1 and conduction electron 'e' due to thermal energy at moderate temperatures.

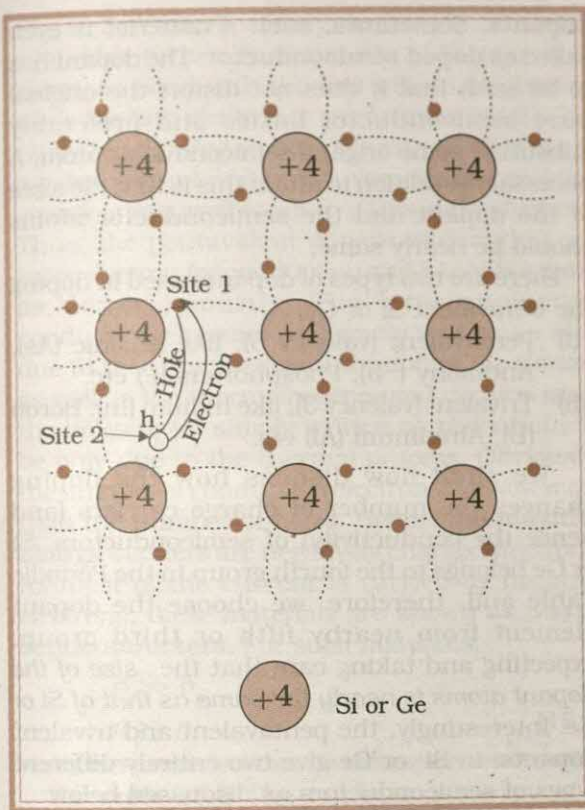


Fig. 15.4(b) Simplified representation of possible thermal motion of a hole 'h'. The electron from the lower left hand covalent bond (site 2) goes to the earlier hole site 1, leaving a hole at its site indicating an apparent movement of the hole from site 1 to site 2.

produced as a result of such an ionisation has been found theoretically to be given by

$$n_e = C T^{3/2} \exp(-E_g/2kT) \quad (15.1)$$

where C is a constant, k is the Boltzmann constant and T is the absolute temperature. For a given E_g , n_e increases as the temperature increases.

Example 15.1 Is the ionisation energy of an isolated free atom different from the ionisation energy E_g for the atoms in a crystalline lattice?

Answer The ionisation energy of an isolated atom is different from its value in crystalline lattice, because in the latter case each bound electron is influenced by many atoms in the periodic crystalline lattice.

Example 15.2 C, Si and Ge have same lattice structure. Why is C insulator while Si and Ge intrinsic semiconductors?

Answer The 4 bonding electrons of C, Si or Ge lie, respectively, in the second, third and fourth orbit. (Fig. 15.1). Hence, energy required to take out an electron from these atoms (i.e. ionisation energy E_g) will be least for Ge, followed by Si and highest for C. Hence, number of free electrons for conduction in Ge and Si are significant but negligibly small for C.

We have seen above that the creation or liberation of **one** free electron by the thermal energy has created **one** hole. Thus in *intrinsic semiconductors*, the number of free electrons (n_e) is equal to the number of holes (n_h). The n_e and n_h are essentially the negative and positive **intrinsic charge carriers** (n_i) such that

$$n_e = n_h = n_i \quad (15.2)$$

Semiconductors possess the unique property in which, apart from electrons, the holes also move. Suppose there is a hole at site 1 as shown in Fig. 15.4(a). The movement of holes can be visualised as shown in Fig. 15.4(b). An electron from the covalent bond at the lower left hand corner (site 2) may jump to the vacant site 1 (hole). Thus, after such a jump, the hole is at site 2 and the site 1 has now an electron. Therefore, apparently, the hole has moved from site 1 to site 2. Thus, the motion of hole may be looked upon as a *transfer of ionisation* from one atom to another carried out by the motion of the bound electrons between their covalent bonds. Note that the electron originally set **free** [see Fig. 15.4(a)] is not involved in this process of hole motion. The **free** electron moves completely independently as conduction electron and gives rise to an electron current (I_e) under an applied electric field. Remember that the motion of hole is only a convenient way of describing the actual motion of **bound** electrons, whenever there is an empty bond anywhere in the crystal. Under the action of an electric field in a real crystal, these holes move towards negative potential (as a result of the sequential jumping of electrons in the reverse direction from one atom to the other) giving the hole current I_h . The total current is the sum of the electron current I_e due to the thermally generated conduction electrons and the hole current I_h .

$$I = I_e + I_h \quad (15.3)$$

15.3.2 Valence Bond Description of Extrinsic Semiconductors

The intrinsic semiconductors discussed earlier in Section 15.3.1 have several limitations when it comes to their use for developing semiconductor based devices. These are:

- The number of intrinsic charge carriers (holes and electrons) is very small $\sim 10^{16} \text{ m}^{-3}$. Hence, these are low conductivity materials.
- The intrinsic charge carriers are always thermally generated and hence enough flexibility is not available to control their number.
- Since $n_e = n_h$ for intrinsic semiconductors, we are forced to accept the situation that the intrinsic materials can never be such that they only have either predominant hole or electron conduction. Again, this limits the usefulness of such materials.

To overcome this problem, a small amount, say, ~ 1 part per million (ppm), of a suitable impurity is added to the pure semiconductor. Such materials are known as *extrinsic semiconductors* or *impurity semiconductors*. The deliberate addition of a desirable impurity is called **doping** and the impurity atoms are called

dopants. Sometimes, such a material is even called as **doped semiconductor**. The dopant has to be such that it does not distort the original pure semiconductor lattice and preferably substitute some original semiconductor atom. A necessary condition to attain this is that the sizes of the dopant and the semiconductor atoms should be nearly same.

There are two types of dopants used in doping the tetravalent Si or Ge:

- Pentavalent (valency 5); like Arsenic (As), Antimony (Sb), Phosphorous (P) etc.
- Trivalent (valency 3); like Indium (In), Boron (B), Aluminum (Al) etc.

We shall now discuss how the doping changes the number of charge carriers (and hence the conductivity) of semiconductors. Si or Ge belongs to the fourth group in the Periodic Table and, therefore, we choose the dopant element from nearby fifth or third group, expecting and taking care that the *size of the dopant atoms is nearly the same as that of Si or Ge*. Interestingly, the pentavalent and trivalent dopants in Si or Ge give two entirely different types of semiconductors as discussed below:

(a) n-type semiconductor

Suppose we dope Si or Ge (valency 4) with a pentavalent (valency 5) element as shown in

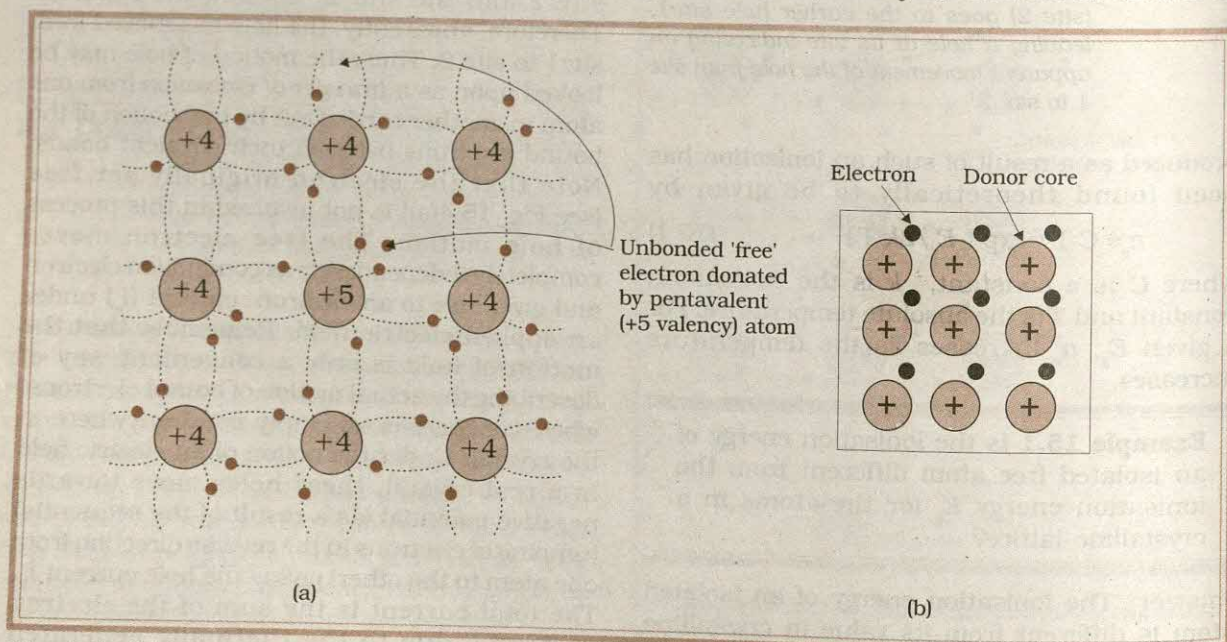


Fig. 15.5 (a) Pentavalent donor atom (As, Sb, P etc.) doped for tetravalent Si or Ge giving n-type semiconductor, and (b) Commonly used schematic representation of n-type material which shows only the fixed cores of the substituent **donors** with one additional effective positive charge and its associated extra electron.

Fig. 15.5. When +5 valency element (As, Sb, P etc.) sits substitutionally at the site of Si, four of its electrons bond with the four silicon neighbours while the fifth electron is free to move (as it is not really functioning to hold the crystal together in a valence bond and only a very small ionisation energy is required to set free this extra electron). Thus, the pentavalent dopant is donating one extra electron for conduction and hence is known as **donor** impurity. The total number of conduction electrons n_e (simply written as n) is due to the combined contribution of the **donors** as well as the thermally generated carriers while the holes n_h (or simply written as p) continue to be only due to the thermal process. Obviously, the number of conduction electrons are now more than the number of holes. Hence, the **majority charge carriers** are negatively charged electrons. As most of the current is carried by negative electrons, these materials are known as **n-type semiconductors**. For such materials,

$$\begin{aligned} n_e &\gg n_h \\ n &\gg p \end{aligned} \quad (15.4)$$

Example 15.3 The ionisation energy of isolated pentavalent phosphorous atom is very large. How is it possible that when it goes into silicon lattice position to release its fifth electron at room temperature so that n-type semiconductor is obtained?

Answer This is quite tricky and follows from the general rule that ionisation energy in lattice is much smaller than that for the isolated atom. Consider Fig. 15.5(a). The +5 valence atom phosphorous is sitting in place of +4 silicon. So it has a *net additional +1 electronic charge*. The four valence electrons form covalent bond and get fixed in the lattice. The fifth electron (*with net -1 electronic charge*) can be **approximated** to revolve around +1 additional charge as explained above. The situation is like the hydrogen atom (Chapter 13) for which energy is approximately given by

$$E_n \approx -13.6/n^2 \text{ eV}$$

$$\text{or } E_1 \approx -13.6 \text{ eV for } n = 1$$

For the case of hydrogen, the permittivity was taken as ϵ_0 . However, if the medium has a permittivity, ϵ_r , relative to ϵ_0 then the above expression can be written as

$$E_1 \approx -13.6/\epsilon_r^2 \text{ eV}$$

For Si, $\epsilon_r = 12$. Hence, E_1 becomes nearly 144 times less $\approx 0.1 \text{ eV}$. So ionisation is possible.

(b) p-type semiconductor

This is obtained when Si or Ge (tetravalent) is doped with group-III trivalent impurities like Al, B, In etc. as shown in Fig. 15.6. The dopant has one outer electron less than Si or Ge and, therefore, this atom can form bonds from three sides with Si and fails to form bond on one side. To hold the dopant atom (In, B or Al) tightly within the silicon or germanium lattice, some of the outer bound electrons in the neighbourhood have a tendency to slide into this vacant bond, as shown in Fig. 15.6, leaving a vacancy or hole at its own site. This **hole** is available for conduction. Note that the trivalent foreign atom (In, B or Al) becomes effectively negatively charged when all its valence bonds are filled. Therefore, many times in common usage, the p-type material is designated as **fixed core of one negative charge** alongwith its associated hole as shown in Fig. 15.6(b). It is obvious that one ionised acceptor atom (N_A) gives one hole. These holes are in addition to the thermally generated holes while the source of conduction electrons is only thermal generation. Thus, for such a material, the holes are the majority carriers and electrons are minority carriers. Obviously, for p-type material

$$n_h \gg n_e \text{ or } p \gg n \quad (15.5)$$

Now, we recapitulate the salient points discussed above. Adding donor (pentavalent) impurity (N_D) in Si or Ge, we create additional conduction electrons, while the dopant atoms themselves become ionised +ve cores. Similarly, additional holes are created by adding acceptors or trivalent impurity (N_A) and N_A atoms themselves become ionised -ve cores. Note that *the crystal maintains an overall charge neutrality*. It may be noted that apart from the **process of generation** of conduction electrons n_e and holes n_h , a simultaneous **process of destruction** occurs in which the electrons **recombine** with the holes. At equilibrium, the **rate of generation of charge carriers is equal to the rate of destruction of charge carriers**.

The recombination occurs due to an electron colliding with a hole. Larger the value of n_e or n_h , higher is the probability of their recombination with each other. Hence, for an extrinsic semiconductor:

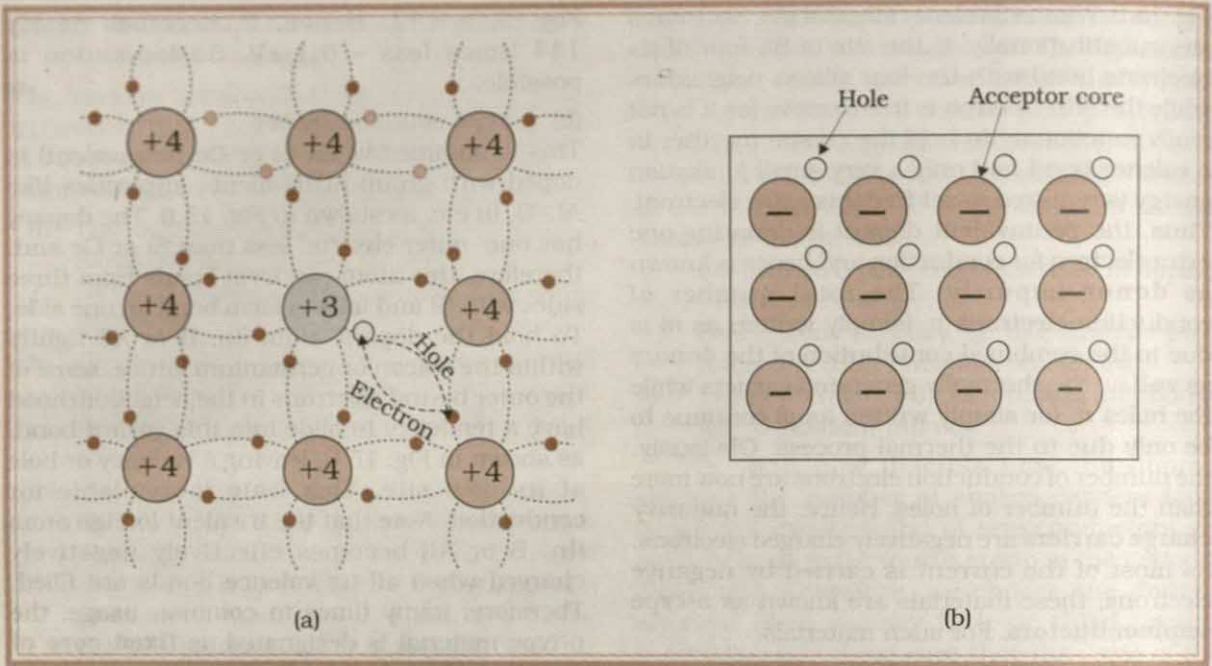


Fig. 15.6 (a) Trivalent acceptor atom (In, Al, B etc.) doped in tetraivalent Si or Ge lattice giving p-type semiconductor. (b) Commonly used schematic representation of p-type material which shows only the fixed core of the substituent **acceptor** with one effective additional negative charge and its associated hole.

Rate of recombination $\propto n_e n_h$

Thus, rate of recombination

$$= R n_e n_h \quad (15.6)$$

where R is a constant known as *recombination coefficient*.

For the special case of intrinsic semiconductor, $n_e = n_h = n_i$ and hence for an intrinsic semiconductor the Eq. (15.6) reduces to:

$$\text{Rate of recombination} = R n_i^2 \quad (15.7)$$

The values of R , the rate of recombination or the rate of generation are governed by laws of thermodynamics, and hence they will remain the same so long as the crystalline lattice structure remains the same. Therefore, the rates of recombination given by Eqs. (15.6) and (15.7) for extrinsic and intrinsic semiconductors are equal, so that

$$R n_e n_h = R n_i^2 \quad (15.8)$$

or $n_e n_h = n_i^2$

This equation is very important in semiconductor physics because of the following implications:

- (i) For n-type semiconductor: n_e is necessarily greater than n_i and yet its product with n_h continues to be n_i^2 [Eq. (15.8)]. This is possible

only if n_h becomes less than n_i . This means that the number of holes get suppressed.

- (ii) For p-type semiconductor: Here n_e would be less than n_i since n_h is necessarily more than n_i . Thus, in p-type material the number of electrons is suppressed to a value lower than even n_i .

Example 15.4 Suppose a pure Si crystal has 5×10^{28} atoms m^{-3} . It is doped by 1 ppm concentration of pentavalent As. Calculate the number of electrons and holes. Given that $n_i = 1.5 \times 10^{16} \text{ m}^{-3}$.

Answer Note that thermally generated electrons ($n_i \sim 10^{16} \text{ m}^{-3}$) are negligibly small as compared to those produced by doping. Therefore, $n_e \approx N_D$. Since $n_e n_h = n_i^2$, The number of holes $n_h = (2.25 \times 10^{32}) / (5 \times 10^{22}) \sim 4.5 \times 10^9 \text{ m}^{-3}$ ◀

15.4 ENERGY BAND DESCRIPTION OF SOLIDS — METALS, INSULATORS AND SEMICONDUCTORS

In the bond description of solids (Section 15.3), the bonding electrons and holes have been considered as highly **localised**. However, this

description is not exact. Due to strong overlapping of the orbitals, it is difficult to ascribe any one electron (or hole) belonging to any particular atom. The problem of electron wave motion is, in fact, quantum mechanical in nature. This viewpoint is more closely related with the **energy and momentum concepts** rather than the **space-localisation and velocity concepts** so far used in the **bond picture**. This alternative approach is termed as **Energy-Band description** of solids and is briefly described in this section. The calculation of the possible electron energies in solids (with $\sim 10^{29}$ atoms per m^3) is very difficult as compared to the case of a single isolated atom as done by you for hydrogen atom in Chapter 13. In fact, it is a quantum mechanical problem to be solved by using Schrodinger's wave equation which is beyond the scope of this book. However, we can easily understand that the electron energies in solids, in view of strong overlap of different atomic orbitals, will be more like an **energy band** instead of discrete energy levels of single isolated atoms.

Consider that the Si or Ge crystal contains N atoms. Electrons of each atom will have discrete energies in different orbits. The electron energy will be same if all the atoms are **isolated**, i.e., separated from each other by a large distance. However, in a crystal, the atoms are close to each other (2 to 3 Å) and therefore the electrons interact with each other and also with the neighbouring atomic cores. The overlap (or interaction) will be more felt by the electrons in the outermost orbit while the inner orbit or core electron energies may remain unaffected. Therefore, for understanding electron energies in Si or Ge crystal, we need to consider the changes in the energies of the electrons in the outermost orbit only. For Si, the outermost orbit is the third orbit ($n = 3$), while for Ge it is the fourth orbit ($n = 4$). The number of electrons in the outermost orbit is 4 (2s and 2p electrons). Hence, the total number of outer electrons in the crystal is $4N$. The maximum possible number of outer electrons in the orbit is 8 (2s + 6p

electrons). So, out of the $4N$ electrons, $2N$ electrons are in the $2N$ **s-states** (orbital quantum number $l = 0$) and $2N$ electrons are in the available $6N$ **p-states**. Obviously, some p-electron states are empty as shown in the extreme right of Fig. 15.7. This is the case of well separated or isolated atoms [region A of Fig. 15.7(a)].

Suppose these atoms start coming nearer to each other to form a solid. The energies of these electrons in the outermost orbit may change (both increase and decrease) due to the interaction between the electrons of different atoms. The $6N$ states for $l = 1$, which originally had identical energies in the isolated atoms, spread out and form an **energy band** [region B in Fig. 15.7(a)]. Similarly, the $2N$ states for $l = 0$, having identical energies in the isolated atoms, split into a second band [carefully see the region B of Fig. 15.7(a)] separated from the first one by an **energy gap**.

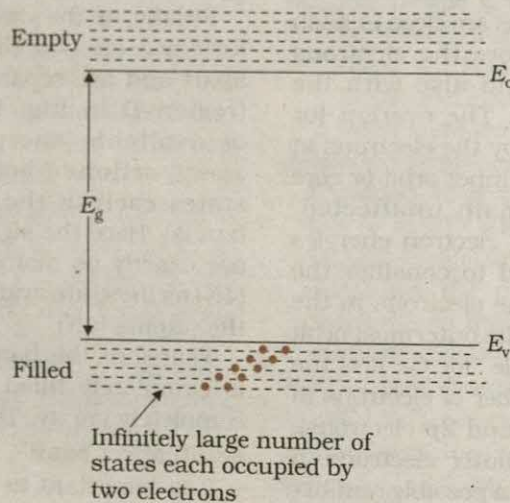
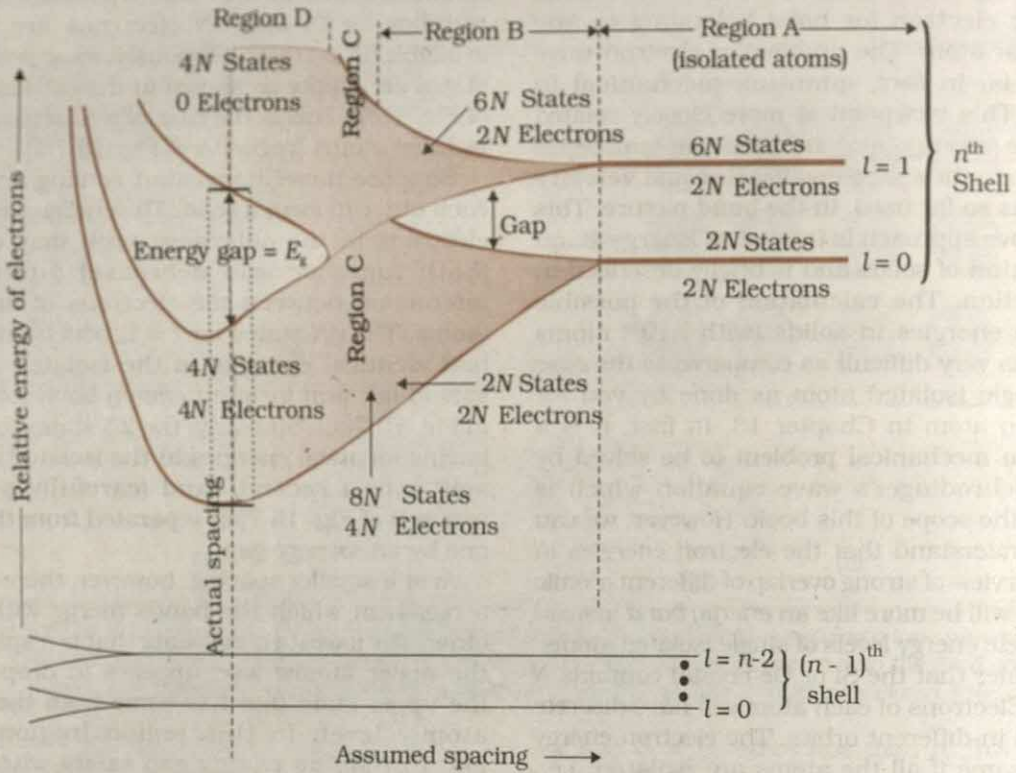
At still smaller spacing, however, there comes a region in which the bands merge with each other. The lowest energy state that is a split from the upper atomic level appears to drop below the upper state that has come from the lower atomic level. In this region [region C in Fig. 15.7(a)], **no energy gap exists where the upper and lower energy states get mixed**.

Finally, if the distance between the atoms further decreases, the energy bands again split apart and are separated by an **energy gap** E_g (region D in Fig. 15.7). The total number of available energy states $8N$ has been **re-apportioned** between the two bands ($4N$ states each in the lower and upper energy bands). Here the significant point is that there are exactly as many states in the lower band ($4N$) as there are available valence electrons from the atoms ($4N$).

Therefore, this band (called the **valence band**) is completely filled while the upper band is completely empty. The upper band is called the **conduction band***.

It is important to realise that at equilibrium spacing, the lowest conduction band energy

* The action of the electrons filling up the **valence band** corresponds to their forming all the **bonds**. However, it is not correct to use the bond description to state that the electrons in the valence band would not move under the action of applied voltage or field. On the basis of quantum mechanics, we know that it is not possible to impart any net momentum to a completely filled band but for the motion of electrons under an electric field in a solid (i.e., electric current) we need the momentum to be given to the electrons in the direction of the field. Therefore, the valence electrons do not conduct. The electrons in the upper band (or conduction band) can, however, gain momentum and move since there are closely spaced empty available states in the band.



(b)

Fig. 15.7 (a) Evolution of energy bands from atomic energy levels as atoms come closer to each other to form a solid. (b) The energy band positions in a real solid. The upper band, called the conduction band, consists of infinitely large number of closely spaced energy states. The lower band, called the valence band, consists of closely spaced completely filled energy states. Note that only two electrons are allowed in each energy state according to Pauli's exclusion principle.

is E_c and highest valence band energy is E_v . Above E_c or below E_v there are a large number of closely spaced energy levels as shown in Fig. 15.7(b). Maximum number of two electrons can be in each energy level due to Pauli's exclusion principle.

The gap between the top of the valence band and bottom of the conduction band is called the energy band gap (Energy gap). It may be large, small, or zero, depending upon the material. These different situations, are depicted in Fig. 15.8 and discussed below:

Case (I): This refers to a situation, as shown in Fig. 15.8(a), where the conduction and valence bands are overlapping. This is the case of a metal where $E_g = 0$. This situation makes a large number of electrons available for electrical conduction and, therefore, the resistance of such materials is low or the conductivity is high.

Case (II): In this case, as shown in Fig. 15.8(b), a large band gap E_g exists ($E_g > 3$ eV). There are no electrons in the conduction band, and therefore no electrical conduction is possible. Note that the energy gap is so large that electrons cannot be easily excited from the valence band to the conduction band by any external stimuli (electrical, thermal or optical). This is the case of **insulators**.

Case (III): This situation is shown in Fig. 15.8(c). Here a finite but small band gap ($E_g < 3$ eV) exists. Because of the small band gap, some electrons can be thermally excited to the **conduction band** (according to Boltzmann law $n \propto \exp(-E_g/2kT)$). These thermally excited electrons (though small in number) can move in the conduction band. Hence, the resistance would not be as high as that of the insulators. This is the case of **semiconductors**.

We have implicitly said above, while describing the semiconductor in terms of band gap, that the conduction band is completely empty in the absence of thermal energy (i.e., $T = 0$ K) as shown in Fig. 15.9(a)(i). Hence, an intrinsic semiconductor will behave like an insulator at $T = 0$ K. It is the thermal energy at high temperatures ($T > 0$ K) which excites some electrons from the valence band (thus creating an equal number of holes in the valence band) to the conduction band. These thermally excited electrons at $T > 0$ K, partially occupy some states in the conduction band. Therefore, the energy-band diagram of an intrinsic semiconductor will

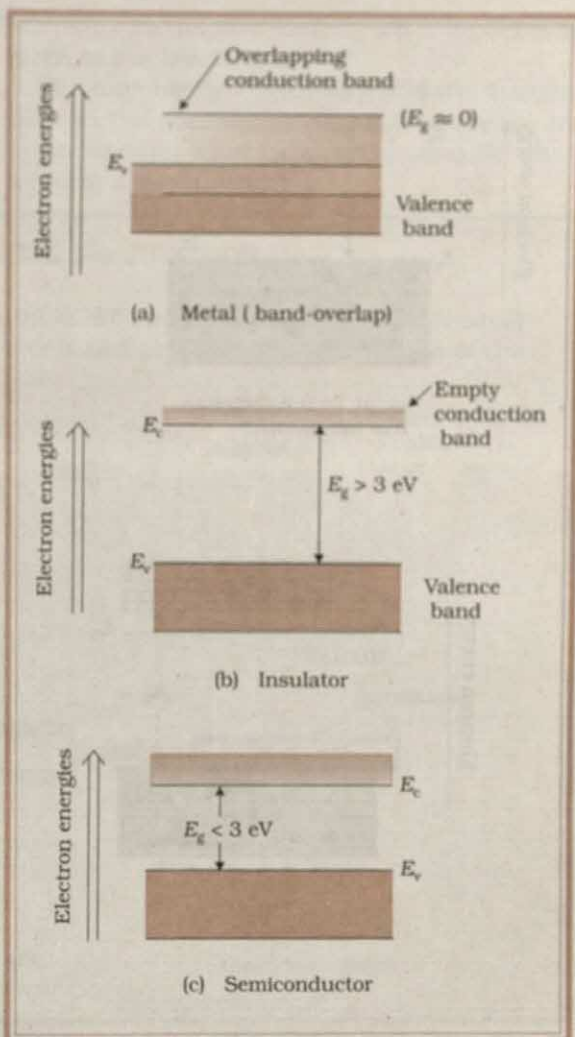


Fig. 15.8 Difference between energy bands of metals, insulators and semiconductors.

be as shown in Fig. 15.9(a)(ii). Here, some electrons are shown in the conduction band which have come from the valence band leaving equal number of holes there. In the case of extrinsic semiconductors, additional energy states, apart from E_c and E_v due to donor impurities (E_d) and acceptor impurities (E_a) also exist. We have discussed in Example 15.3 of Section 15.3.2 that very small energy (~ 0.1 eV) is required for the electrons to be released from the donor impurity in the n-type semiconductor. Hence the donor energy level E_d lies very near the bottom of the conduction band which can pump electrons into the conduction band. The conduction band will now have more electrons as they have come from **thermal excitation** as

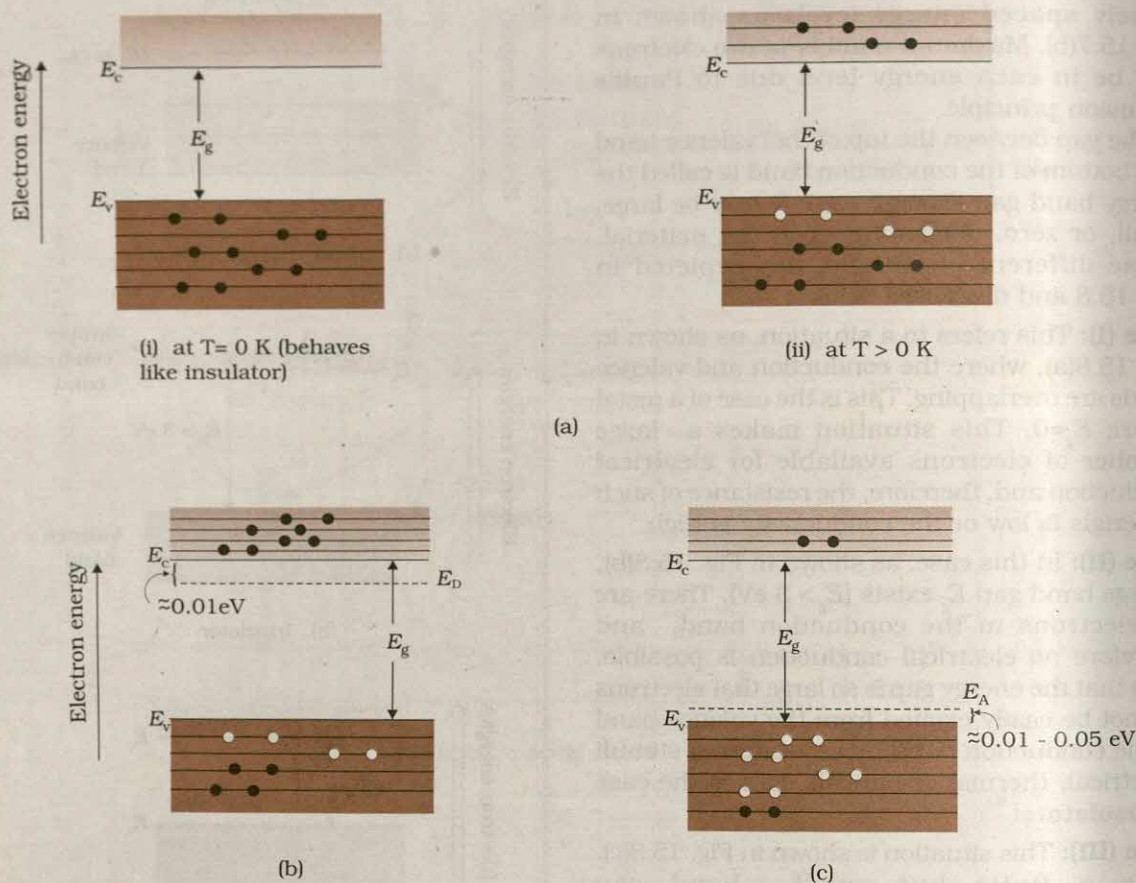


Fig. 15.9

Fig. 15.9 Energy bands in different types of semiconductors. The filled circle symbol (\bullet) is for electrons and empty circle (\circ) is for holes: (a) Intrinsic semiconductor; (b) n-type semiconductor ($T > 0$ K); and (c) p-type semiconductor ($T > 0$ K).

well as from the donor impurities as shown in Fig. 15.9(b). Similarly, for p-type semiconductor, an acceptor energy level E_A is obtained as shown in Fig. 15.9(c). The position of E_A is very near to the top of the valence band because, as we have argued in Section 15.3.2, an electron added to the acceptor impurity to complete its bonding within the semiconductor structure, comes easily from the valence band electrons of some other semiconductor atoms in the lattice.

The representation given in Figs. 15.8 and 15.9 are referred to as **Energy-Band diagrams**. Though the above description is grossly approximate and hypothetical, it helps in understanding the difference between metals, insulators and semiconductors (extrinsic and

intrinsic) in a simple manner. The energy-band diagrams have been worked out precisely by quantum mechanical methods for a crystal of finite size. This is beyond the scope of this book.

15.5 p-n JUNCTION — BASIC UNIT OF ALL SEMICONDUCTOR DEVICES

A p-n junction is at the core of almost all semiconductor devices. For example, a single p-n junction acts as a rectifying diode (explained later in Sections 15.6 and 15.7). Similarly, in a p-n-p transistor there are two such junctions (viz. p-n followed by n-p). Hence, it is important that we clearly understand the p-n junction before going into the details of various other devices.

The p- and n-type silicon or germanium can be obtained by adding appropriate acceptor or donor impurity into Si- or Ge- melt while growing a crystal. These crystals are cut into thin slices called *wafers*. Semiconductor devices are usually made on these wafers. There

are many methods of making a p-n junction as given in the box below.

The importance of p-n junction will become clear in the subsequent sections where we try to understand what happens electrically when we form a p-n junction.

TECHNIQUES FOR MAKING p-n JUNCTION

(i) **Alloy junction:** This is shown in Fig.15.10(a). We start with n-Ge or n-Si. A small piece of III-group metal Indium is placed over it and melted. The lower portion of the

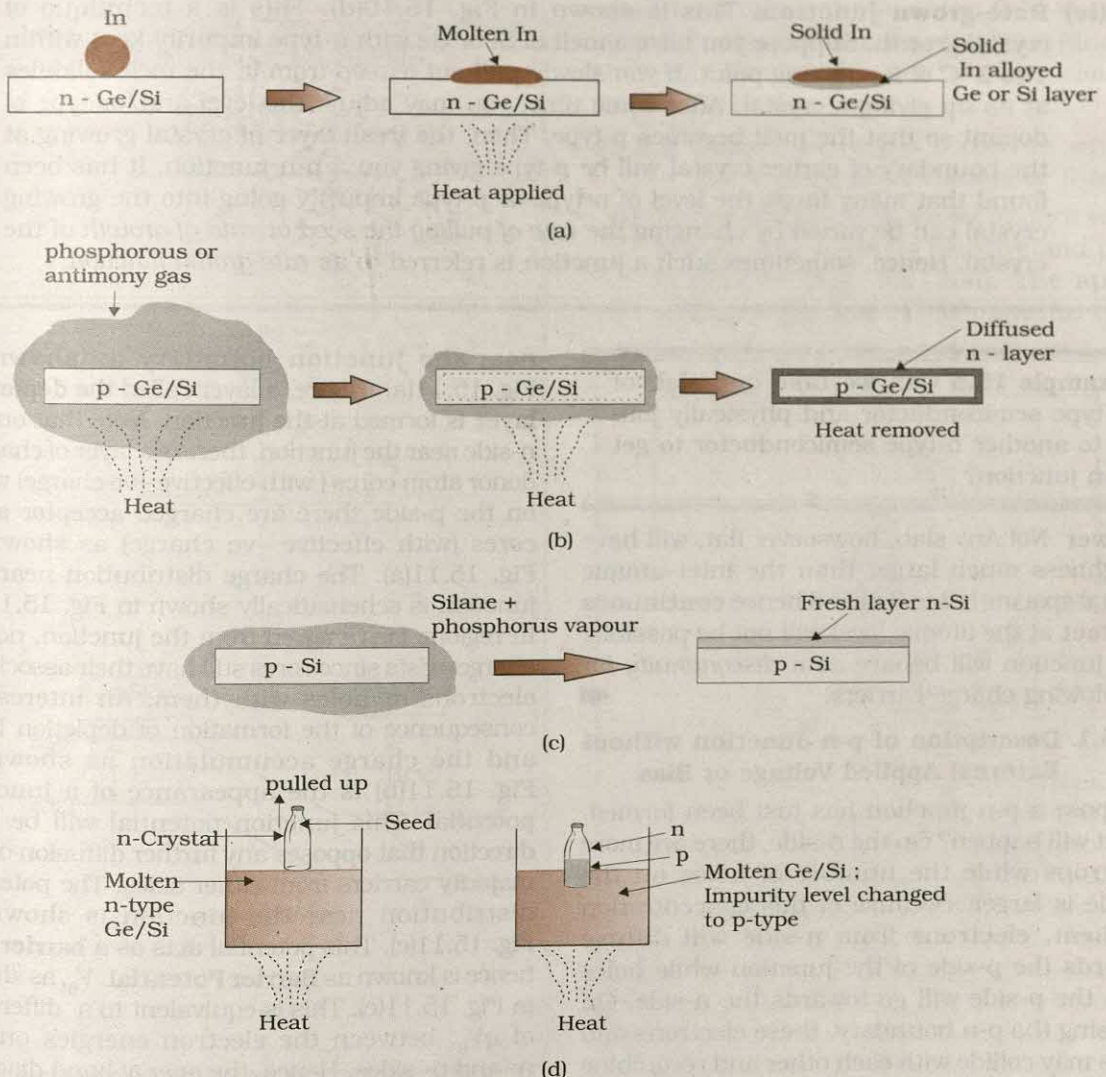


Fig. 15.10

Fig. 15.10 Different techniques of forming a p-n junction. (a) 'Alloying' technique for p-n junction; (b) Diffusion technique; (c) Vapour deposition (epitaxial growth); and (d) Rate-grown junction.

molten indium forms **alloy** with the n-semiconductor and converts its top layer into p layer giving p-n junction.

- (ii) **Diffusion junction:** This is shown in Fig. 15.10(b). A heated p-type semiconductor is kept in phosphorous or antimony vapours which diffuse into the p-semiconductor making the top layer as n-type (since P or Sb are V-group donor impurity).
- (iii) **Vapour deposited junction:** It is shown in Fig. 15.10(c). Suppose you have to grow a layer of n-Si on p-Si. The p-Si wafer is kept in an atmosphere of Silane (a silicon compound which dissociates into Si at high temperatures) plus phosphorous vapours. On cracking of Silane at high temperatures, a fresh layer of n-Si grows on p-Si giving the p-n junction. Since this junction growth is *layer-by-layer*, it is also referred to as *layer grown junction* (or epitaxial junction).
- (iv) **Rate-grown junction:** This is shown in Fig. 15.10(d). This is a technique of crystal growth. Suppose you have a melt of Si or Ge with n-type impurity kept within 1 to 2 °C of its melting point. If you slowly pull out a *seed* from it, the melt solidifies at its tip giving a crystal. After some time, you may adjust the level and/or type of dopant so that the melt becomes p-type. Then, the fresh layer of crystal growing at the boundary of earlier crystal will be p-type giving you a p-n junction. It has been found that many times the level of n-type or p-type impurity going into the growing crystal can be varied by changing the *rate of pulling the seed* or *rate of growth* of the crystal. Hence, sometimes such a junction is referred to as *rate-grown junction*.

Example 15.5 Can we take one slab of p-type semiconductor and physically join it to another n-type semiconductor to get p-n junction?

Answer No! Any slab, howsoever flat, will have roughness much larger than the inter-atomic crystal spacing (~ 2 to 3 \AA) and hence **continuous contact** at the atomic level will not be possible. The junction will behave as a *discontinuity* for the flowing charge carriers. ◀

15.5.1 Description of p-n Junction without External Applied Voltage or Bias

Suppose a p-n junction has just been formed. What will happen? On the n-side, there are more electrons while the number of holes on the p-side is larger. Because of this concentration gradient, electrons from n-side will diffuse towards the p-side of the junction while holes from the p-side will go towards the n-side. On crossing the p-n boundary, these electrons and holes may collide with each other and recombine (or annihilate) since they have opposite charges. These electrons/holes have come from donor or acceptor impurity atom cores. Hence, such donor or acceptor atoms will get *depleted* of their associated electrons or holes and subsequently will be left with a **charged ion core** in the layer

near the junction boundary as shown in Fig. 15.11(a). Hence, a layer called the **depletion layer** is formed at the junction. Note that on the n-side near the junction, there is a layer of charged donor atom cores (with effective +ve charge) while on the p-side there are charged acceptor atom cores (with effective -ve charge) as shown in Fig. 15.11(a). The charge distribution near the junction is schematically shown in Fig. 15.11(b). In regions far removed from the junction, no net charge exists since cores still have their associated electrons or holes with them. An interesting consequence of the formation of depletion layer and the charge accumulation as shown in Fig. 15.11(b) is the appearance of a junction potential. This junction potential will be in a direction that opposes any further diffusion of the majority carriers from either sides. The potential distribution near the junction is shown in Fig. 15.11(c). This potential acts as a **barrier** and hence is known as **Barrier Potential**, V_B , as shown in Fig. 15.11(c). This is equivalent to a difference of qV_B , between the electron energies on the n- and p- sides. Hence, the energy band diagram for a p-n junction can be drawn as shown in Fig. 15.11(d). You can see that the energy barrier across the junction is now qV_B (it will be $-eV_B$ for electrons as $q = -e$). Note that the energy difference qV_B has to be surmounted before any charge carrier can flow across the junction.

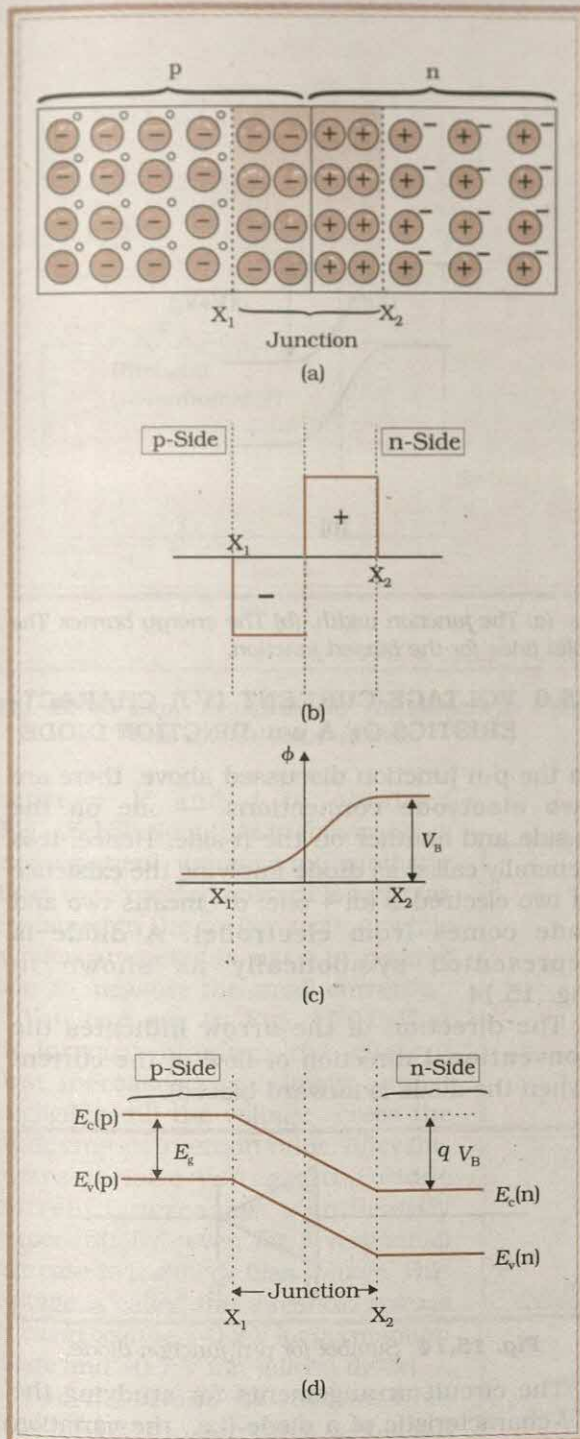


Fig. 15.11 Description of various situations near p-n junction: (a) Formation of depletion layer consisting of immobile negative ion core - and positive ion core +. (b) charge accumulation (c) junction potential (d) energy band positions.

An important point to remember is that the potential barrier (or field across the junction) and the depletion layer width (or junction width) depends upon the doping concentration on the two sides. Suppose N_A and N_D are small. The diffusing electrons and holes across the junction can move to reasonably large distances before suffering a collision with another hole or electron to be annihilated or recombined. Hence, junction width would be large. Obviously, the junction field would be weak. On the other hand, if N_A and N_D are large the junction width would be small (and hence the junction field would be strong). In this manner, we can obtain junctions showing different behaviour by simply changing the doping levels.

15.5.2 Behaviour of p-n Junction with an External Applied Voltage or Bias

- (i) **Forward bias:** Suppose we apply a voltage V such that n-side is negative and p-side is positive [Fig. 15.12(a)]. The applied voltage V (or bias V) is opposite to the

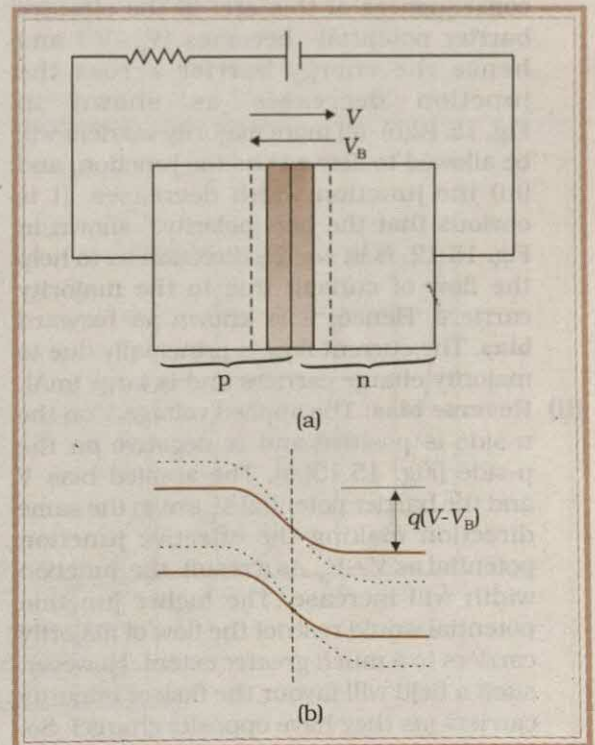


Fig. 15.12 Behaviour of p-n junction under forward bias: (a) Junction width, and (b) energy barrier. The dashed lines are for unbiased junction and solid lines for forward biased junction.

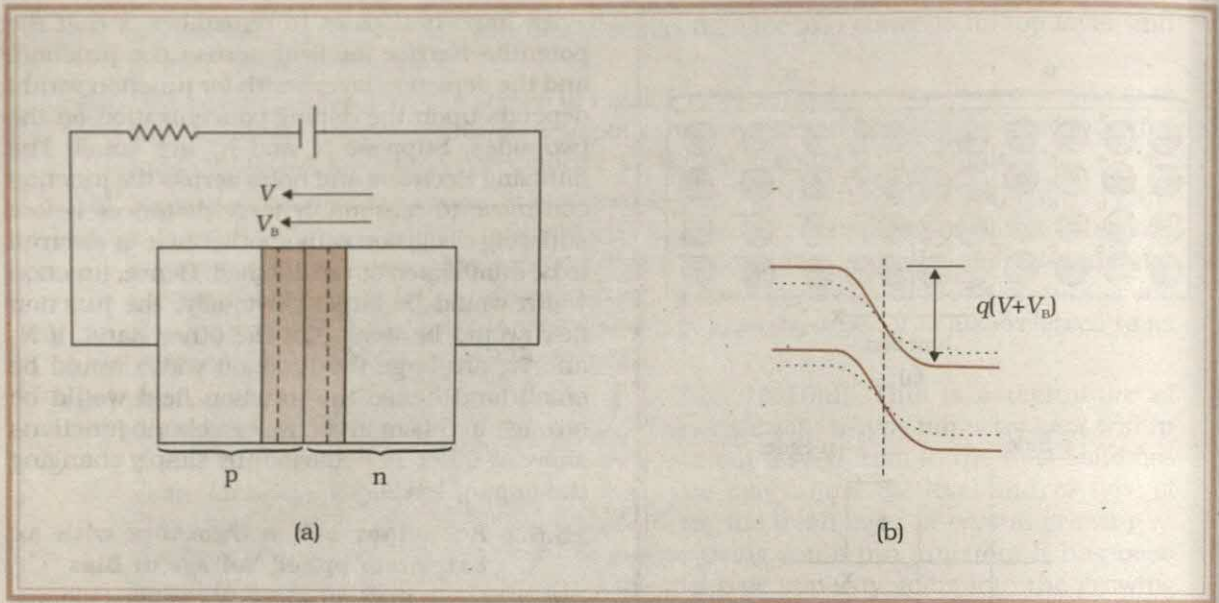


Fig. 15.13 Behaviour of a p-n junction under reverse bias. (a) The junction width. (b) The energy barrier. The dashed lines are for unbiased junction and solid lines for the biased junction.

junction barrier potential. The consequences of this are: (i) the effective barrier potential becomes $(V_B - V)$ and hence the energy barrier across the junction decreases as shown in Fig. 15.12(b), (ii) more majority carriers will be allowed to flow across the junction, and (iii) the junction width decreases. It is obvious that the bias polarity, shown in Fig. 15.12, is in such a direction as to help the flow of current due to the majority carriers. Hence, it is known as **forward bias**. The current flow is principally due to majority charge carriers and is large (mA).

- (ii) **Reverse bias:** The applied voltage V on the n-side is positive and is negative on the p-side [Fig. 15.13(a)]. The applied bias V and the barrier potential V_B are in the same direction making the effective junction potential as $V + V_B$. As a result, the junction width will increase. The higher junction potential would restrict the flow of majority carriers to a much greater extent. However, such a field will favour the flow of minority carriers (as they have opposite charge). So, the reverse bias current will be due to the minority carriers only. Since the number of minority carriers is very small as compared to the majority carriers, the reverse bias current is small ($\sim \mu\text{A}$).

15.6 VOLTAGE-CURRENT (V-I) CHARACTERISTICS OF A p-n JUNCTION DIODE

In the p-n junction discussed above, there are **two electrode** connections — one on the p-side and another on the n-side. Hence, it is generally called as **diode** implying the existence of two electrodes (di + ode; di- means two and -ode comes from electrode). A diode is represented symbolically as shown in Fig. 15.14.

The direction of the arrow indicates the **conventional** direction of flow of the current (when the diode is forward biased).

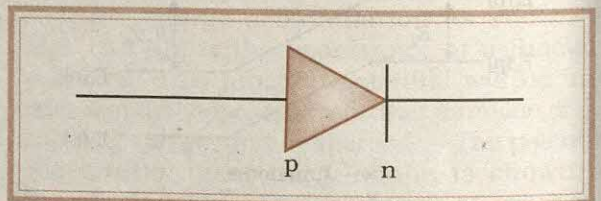


Fig. 15.14 Symbol for p-n junction diode.

The circuit arrangements for studying the V-I characteristic of a diode (i.e., the variation of current as a function of applied voltage) are shown in Fig. 15.15. The voltage is connected to the diode through a potentiometer (or Rheostat) so that the voltage applied to the diode can be changed. For different values of voltages, the value of the current is noted. A graph

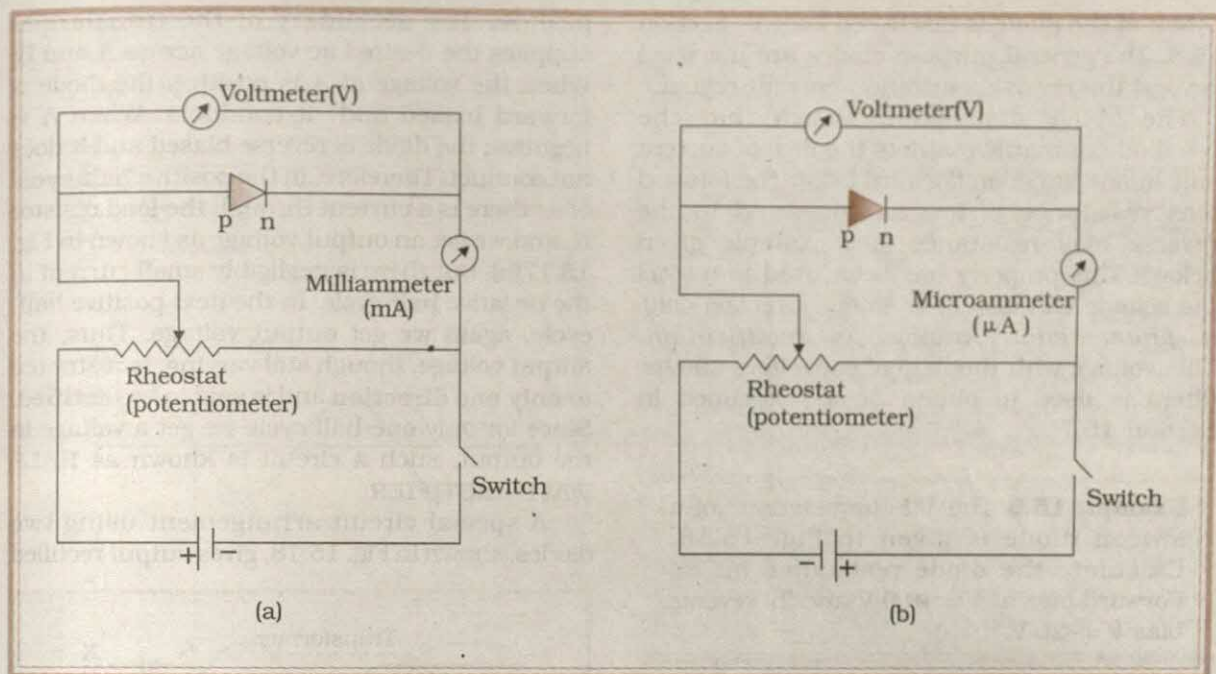


Fig. 15.15 Experimental circuit arrangement for studying V-I characteristics of a p-n junction diode. (a) Forward bias and (b) Reverse bias.

between V and I is obtained (Fig. 15.16). Note that in forward bias measurement, we use a milliammeter since the expected current is large (as explained in the earlier section) while a microammeter is used in reverse bias to measure the small currents.

You can see in Fig. 15.16 that in forward biasing, the current first increases very slowly almost negligibly, till the voltage across the diode crosses a certain value. After this characteristic voltage, the diode current increases significantly (exponentially), even for a very small increase in the diode bias voltage. This voltage is called the *threshold voltage* or *cut-in voltage*, (~ 0.2 V for germanium diode and ~ 0.7 V for silicon diode).

For the diode in reverse bias (Fig. 15.16), the current is very small ($\sim \mu A$) and almost remains constant with bias. It is called **reverse saturation current**. However for special cases, at very high reverse bias (breakdown voltage) the current suddenly increases. This special

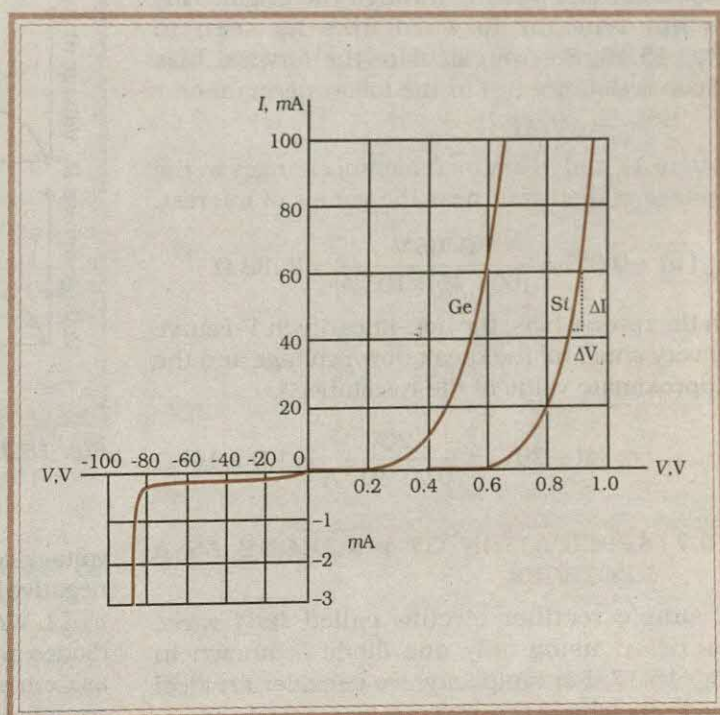


Fig. 15.16 Typical V-I characteristics of silicon and germanium diodes.

action of the diode is discussed later in Section 15.8. The general purpose diodes are not used beyond the reverse saturation current region.

The above discussion shows that the p-n diode primarily restricts the flow of current only in one direction (forward bias). The forward bias resistance is low as compared to the reverse bias resistance (see example given below). This property has been used to restrict the voltage variation of ac to one direction only, a phenomenon known as **rectification**. This voltage with the help of capacitors and/or filters is used to obtain dc as explained in Section 15.7.

Example 15.6 The V - I characteristic of a silicon diode is given in Fig. 15.16. Calculate the diode resistance in: (a) Forward bias at $V = +0.9$ V and (b) reverse bias $V = -20$ V.

Answer The method of calculating resistance from Ohm's law (Chapter 3) by calculating V/I is not strictly valid for the forward bias case of a diode. The Ohm's law assumes the V - I plot to be a straight line passing through the origin. This is not true for forward bias as seen in Fig. 15.16. So, we calculate the forward bias diode resistance (r_{fb}) in the following manner:

$$r_{fb} = \Delta V / \Delta I$$

where ΔV and ΔI are incremental changes in the voltage and current near the values of interest.

$$r_{fb} \text{ (at } +0.9\text{V)} = \frac{0.05 \text{ V}}{(60 - 45) \times 10^{-3} \text{ A}} = 3.33 \Omega$$

In the reverse bias, the non-linearity in V - I curve is very small till the break down voltage and the approximate value of the resistance is

$$r_{rb} \text{ (at } -20\text{V)} = \frac{20\text{V}}{0.2 \times 10^{-3} \text{ A}} = 1.0 \times 10^5 \Omega$$

15.7 APPLICATION OF p-n DIODE AS A RECTIFIER

A simple rectifier circuit, called **half wave rectifier**, using only one diode is shown in Fig. 15.17. For simplicity, we consider an ideal diode in which the reverse bias resistance is

infinite. The secondary of the transformer supplies the desired ac voltage across A and B. When the voltage at A is positive, the diode is forward biased and it conducts. When A is negative, the diode is reverse biased and it does not conduct. Therefore, in the positive **half-cycle** of ac there is a current through the load resistor R_L and we get an output voltage as shown in Fig. 15.17(b), but there is negligibly small current in the negative half-cycle. In the next positive half-cycle, again we get output voltage. Thus, the output voltage, though still varying, is restricted to **only one direction** and is said to be **rectified**. Since for only one-half cycle we get a voltage in the output, such a circuit is known as **HALF WAVE RECTIFIER**.

A special circuit arrangement using two diodes, shown in Fig. 15.18, gives output rectified

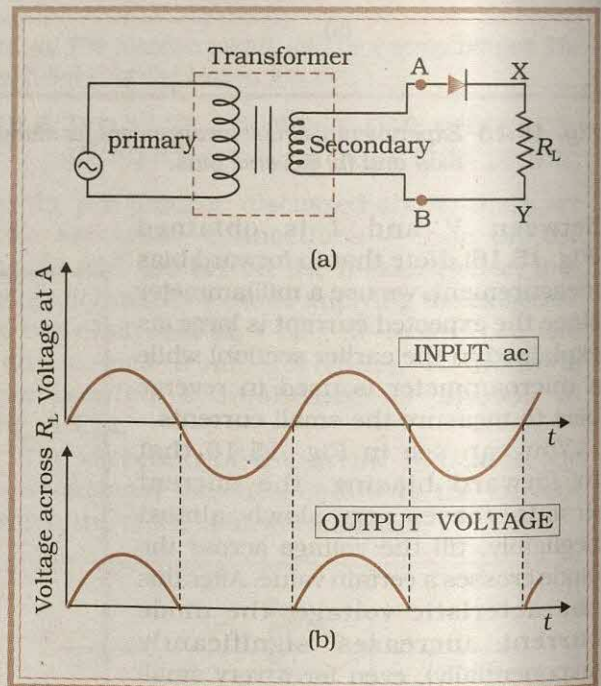


Fig. 15.17 (a) Half-Wave Rectifier circuit. (b) Input ac and output voltage waveforms from the rectifier circuit.

voltage corresponding to the positive as well as negative half of the ac cycle. Hence, it is known as **FULL WAVE RECTIFIER**. The circuit uses two diodes and a special type of transformer known as **Center-tap transformer**: The secondary of the transformer is wound into two equal parts as

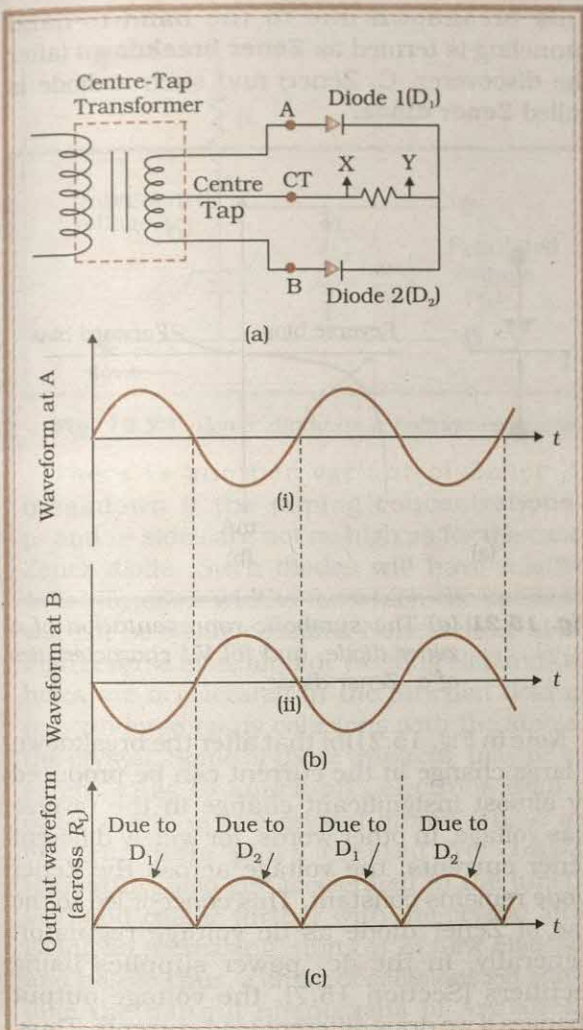


Fig. 15.18 (a) A Full Wave Rectifier circuit; (b) Input wave forms given to the diode D_1 at A and to the diode D_2 at B; (c) Output wave form across the load R_L connected in the Full Wave Rectifier circuit.

shown in the figure. Note that the voltages at any instant at 'A' (input of diode D_1) and B (input of diode D_2) with respect to the center tap are out of phase with each other. Suppose the input voltage to A at any instant is positive. It is clear that, at that instant, voltage at B being out of phase will be negative as shown in Fig. 15.18(b). The diode D_1 gets forward biased and conducts (while D_2 is not conducting). Hence, during this positive half cycle we get an output current (and a consequent output voltage across the load resistor R_L) as shown in Fig. 15.18(c). At another instant, when the voltage at A becomes negative then the voltage at B would

be +ve. Hence, the diode D_1 does not conduct but the diode D_2 conducts giving an output current and output voltage (across R_L) during the negative half cycle of the input ac. Thus, we get output voltage during the +ve as well as the -ve half of the cycle (or in other words, during the full wave). Therefore, such a circuit is known as **FULL WAVE RECTIFIER**. Obviously, this is a more efficient circuit for getting rectified voltage or current.

Note that the rectifier circuits result in the flow of load current (or voltage across R_L) restricted only in one direction unlike the general ac input voltage where the voltage can be both positive and negative. The rectified voltage is still *varying voltage but restricted to only one direction*. You know that all the varying signals can be considered as the sum of a dc signal superimposed with many ac signals of different harmonic frequencies. So, this would also be the case for the rectified voltages. From this rectified voltage, you can obtain dc voltage (or dc power supply) by *filtering out* the ac components. The detailed theory of different filter circuits is beyond the scope of this book. However, you can easily see that even a single capacitor C as shown in Fig. 15.19 filters out the ac component. In general, a high value of C provides a low impedance path to ac but high, almost infinite, impedance to dc (note that the impedance due to C is $1/\omega C$ as discussed in Chapter 8). Hence, ac is bypassed through C or filtered. A predominantly dc like voltage appears at the load.

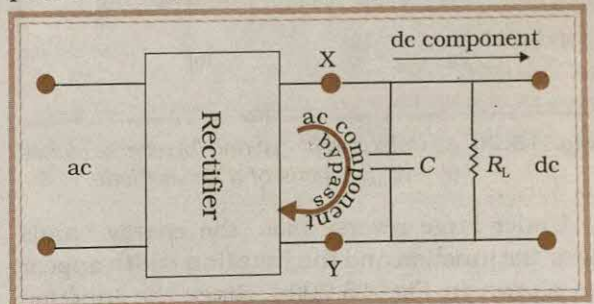


Fig. 15.19 A simple capacitor acting as a *filter* for ac to give dc output voltage from rectified voltages.

15.8 SPECIAL PURPOSE p-n JUNCTION DIODES

15.8.1 Breakdown Diodes (viz. Zener Diode, Avalanche Diode)

We have discussed the forward and reverse biased characteristics of a normal diode earlier

in Section 15.5. In reverse bias, the current is very small and nearly constant with bias (termed as reverse saturation current). However, interesting behaviour results in some special cases if the reverse bias is increased further beyond a certain limit. This is discussed below.

Consider a p-n junction shown in Fig. 15.20(a) where both p- and n-sides are heavily doped. Here, the symbols p^+ and n^+ mean that the semiconductors are heavily doped by acceptor and donor impurities, respectively. Due to the high dopant densities, the depletion layer junction width is small and the junction field will be high (see discussion in Section 15.5.1).

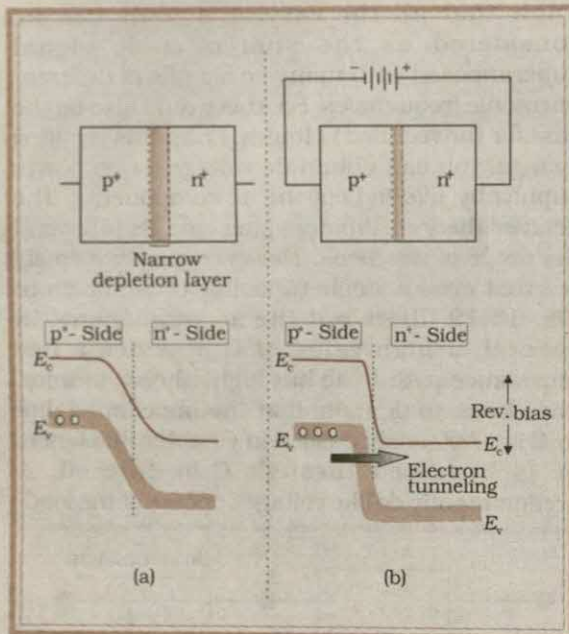


Fig. 15.20 (a) Unbiased $p^+ - n^+$ and (b) reverse biased $p^+ - n^+$ junctions of a Zener diode.

Under large reverse bias, the energy bands near the junction and the **junction width** appear as shown in Fig. 15.20(b). Since the junction width is $< 10^{-7}$ m, even a small voltage (say 4 V) may give a field as large as 4×10^7 Vm⁻¹. The high junction field may strip an electron from the valence band which can **tunnel** to the n-side through the **thin** depletion layer. Such a mechanism of emission of electrons after a certain critical field or applied voltage V_z is termed as **internal field emission** which gives rise to a high reverse current or breakdown current as shown in Fig. 15.21(b).

This breakdown due to the band-to-band tunneling is termed as **Zener breakdown** (after the discoverer, C. Zener) and such a diode is called **Zener diode**.

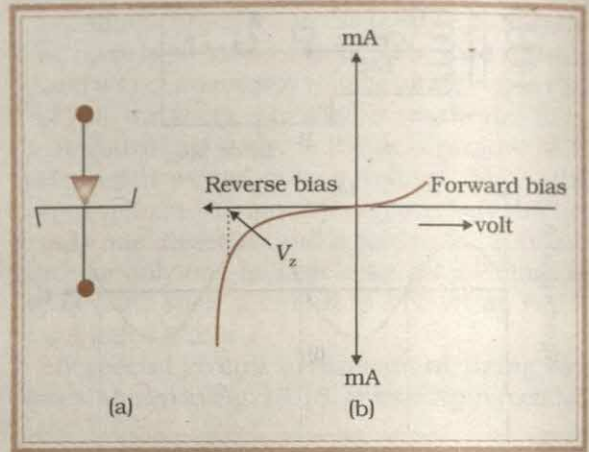


Fig. 15.21 (a) The symbolic representation of a zener diode, and (b) V-I characteristics of a Zener diode.

Note in Fig. 15.21(b) that after the breakdown a large change in the current can be produced by almost insignificant change in the reverse bias voltage. In other words, for widely different Zener currents, the voltage across the Zener diode remains constant. This concept led to the use of Zener diode as dc voltage regulator. Generally, in the dc power supplies using rectifiers (Section 15.7), the voltage output changes if we draw different load currents. Thus, the voltages are **unregulated**. If the output voltage of a dc power supply does not change with load, it is called **regulated power supply**. Suppose an unregulated dc input voltage (V_i) is applied to the Zener diode (whose breakdown voltage is V_z) as shown in Fig. 15.23. If the applied voltage $V_i > V_z$, then the Zener diode is in the breakdown condition. As a result, for a wide range of values of load (R_L) the current in the circuit or through the Zener diode may change but the voltage across it remains unaffected by load as is clear from the Fig. 15.22. Thus, the output voltage across the Zener diode is a **regulated voltage**.

Zener diodes with different breakdown voltages (for regulations of different voltages) can be obtained by changing the doping concentration on its p- and n- sides since they would change junction width (and junction field).

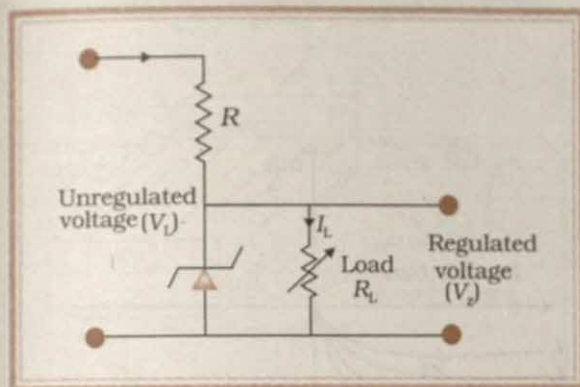


Fig. 15.22 Zener diode as a voltage regulator.

There is another variant of Zener like breakdown if the doping concentrations of p- and n- sides are not as high as for the case of Zener diode. Such diodes will have relatively wider junction widths for which the **tunneling** will not be easily possible. Nonetheless, at very high reverse bias, already existing electrons and holes are accelerated in the junction field and may undergo many collisions with the atoms in the crystal. Some of these **impacts** may be very severe so as to **knock-off** electrons from the outer orbits of the crystal atoms. These **new** electron-hole pairs created by **impact ionisation** also get accelerated in the junction field and collide further with the crystal atoms giving an increasing number of **new electrons and holes**. Thus, a chain of collisions get started (like the natural phenomena of **Avalanche**). This gives rise to very much enhanced number of charge carriers leading to a rapid increase in the junction current at reverse bias beyond a certain critical value. This phenomenon is known as **Avalanche breakdown** and the device is referred to as **Avalanche diode**.

15.8.2 Photonic p-n Junction Devices (Photodiode, Solar Cell, Light Emitting Diode, Diode Laser)

We have seen so far that when a voltage is applied across the p-n junction, it results in a change in current due to the **electron excitation** from the valence band to the conduction band. Both the phenomena of the electron excitation as well as the resulting diode current are **electrical** in nature. However, we can have semiconductor electronic devices, in which the light **photons** have also a role to play in the overall performance of the device. Such devices

are called **photonic** or **opto-electronic devices** which can be classified as:

- (i) Photo-detectors for detecting optical signals (e.g., photodiodes and photoconducting cell).
- (ii) Photovoltaic devices for converting optical radiation into electricity (e.g., solar cells).
- (iii) Devices for converting electrical energy into light (e.g., light emitting diodes and diode lasers)

We discuss next the basic principles of these devices.

(a) Photodiodes: Photodiode is a special type of photo-detector. The general principle of all semiconductor-based photodetectors is the electron-excitation from the valence band to the conduction band by photons. Suppose an optical photon of frequency ν is incident on a semiconductor, such that its energy is greater than the band gap of the semiconductor (i.e., $h\nu > E_g$). This photon will excite an electron from the valence band to the conduction band leaving a vacancy or hole in the valence band. Thus, an electron-hole pair is generated. These are additional charge carriers termed as **photogenerated charge carriers** which obviously increase the **conductivity** of the semiconductor. Larger the number of incident photons (incident intensity of light), larger would be the change in the conductivity of the semiconductor. Therefore, by measuring the change in the conductance (or resistance) of the semiconductor, one can measure the intensity of the optical signal. Such photodetectors are known as **photoconductive cells**. However, more commonly used photo detecting devices are **photodiodes**.

Simplest photodiode is a reverse biased p-n junction diode as shown in Fig. 15.23. You already know that at the p-n junction there exists a **junction field** which, at equilibrium, does not permit the flow of charge carriers across the junction (Section 15.5). The current can only flow with applied bias. The forward bias current is due to the majority carriers while the **reverse bias current** (which is very small) is due to the minority carriers. The diodes are generally reverse biased when used as photodiode (Example 15.7). Suppose such a p-n diode is illuminated with light photons having energy $h\nu > E_g$, and intensities I_1, I_2, I_3 etc. The electron

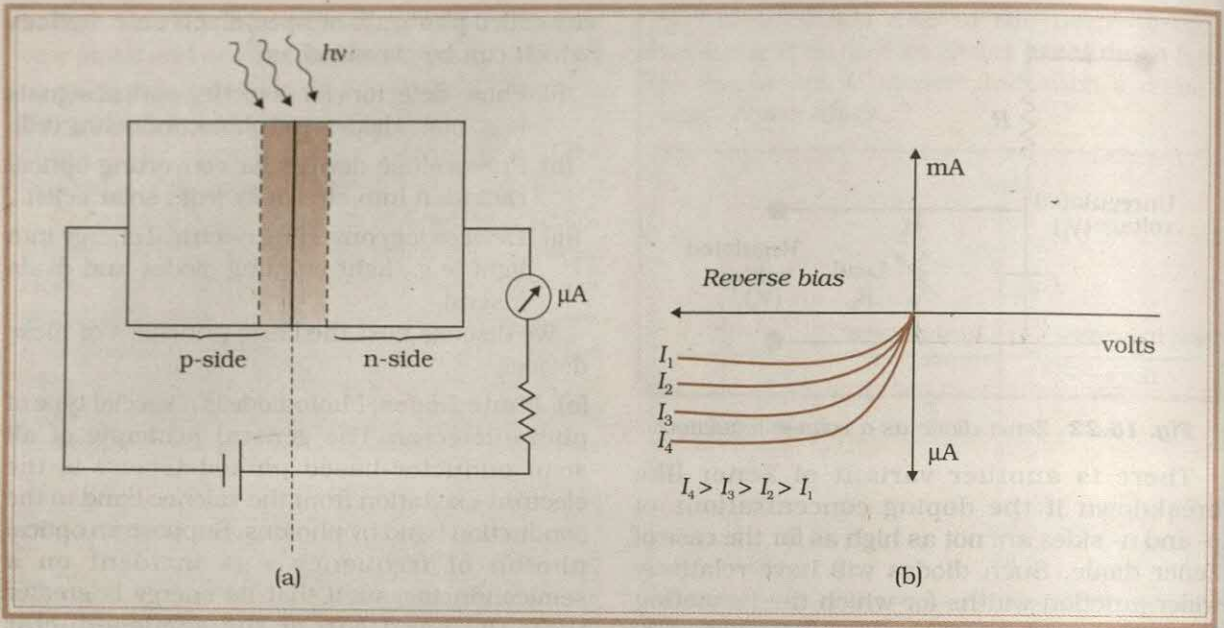


Fig. 15.23 (a) An illuminated photodiode in reverse bias and (b) reverse bias currents under different illumination intensities I_1, I_2, I_3 etc., $I_4 > I_3 > I_2 > I_1$.

and hole pairs generated in the depletion layer (or near the junction) will be separated by the junction field and made to flow across the junction. There would be a change in the reverse saturation current as shown in Fig. 15.23(b). Hence, a measurement of the change in the reverse saturation current on illumination can give the values of the light intensity.

Apart from the p-n junction diode, there are many other diodes like p-i-n diodes (i stands for the insulating layer); Metal-semiconductor diode; p-n avalanche photodiode etc., which are also used as photodetectors. A discussion of these is beyond the scope of this book.

Example 15.7 The current in the forward bias is known to be more (\sim mA) than the current in the reverse bias (\sim μA). What is the reason, then, to operate the photodiodes in reverse bias?

Answer Consider the case of an n-type semiconductor. Obviously, the majority carrier density (n) is considerably larger than the minority hole density p (i.e., $n \gg p$). On illumination, the number of both types of carriers would equally increase as:

$$n^* = n + \Delta n$$

$$p^* = p + \Delta p$$

Remember $\Delta n = \Delta p$ and $n \gg p$. Hence, the fractional change in the majority carriers (i.e., $\Delta n/n$) would be much less than that in the minority carriers (i.e., $\Delta p/p$). In general, we can state that the fractional change due to the photo-effects on the **minority carrier dominated reverse bias current** is more easily measurable than the fractional change in the forward bias current. Hence, photodiodes are preferably used in the reverse bias condition for measuring light intensity.

(b) Solar Cell: Solar cell is a device for converting solar energy into electricity. It is based on similar principle as junction photodiode except that: (i) it is operated in the photovoltaic mode (generation of voltage due to the bombardment of optical photons), (ii) no external bias is applied, and (iii) the active junction area is kept large because we are interested in more power. Different types of solar cells can be obtained by using different types of junctions like p-n or Metal-Semiconductor (M-S) junction or Metal-Oxide Semiconductor (MOS) or Metal-insulator-Semiconductor (M-I-S) junction.

Here, we shall discuss a simple p-n junction solar cell shown in Fig. 15.24. An n-type semiconductor substrate (backed with a current collecting metal electrode) is taken, over which a thin p-layer is grown (e.g., by diffusion of a

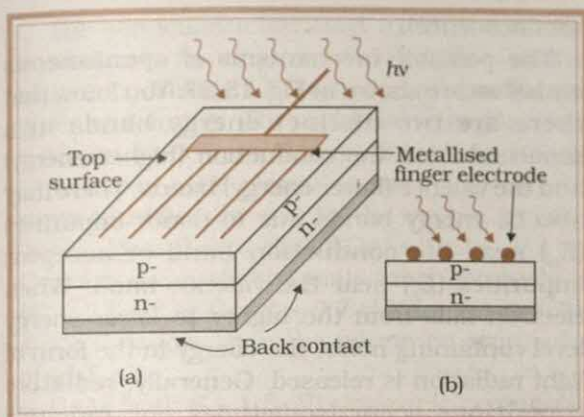


Fig. 15.24 (a) A typical p-n junction solar cell
(b) Sectional view of the solar cell.

suitable acceptor impurity or by vapour deposition). On top of the p-layer, metal **finger** electrodes are prepared so that there is enough space between the fingers for the light to reach p-layer (and the underlying p-n junction) to be able to produce photo-generated holes and electrons. The generation of photo-voltage in the solar cell can be easily understood with the help of Fig. 15.25. Consider the p-n junction as shown in Fig. 15.25(a).

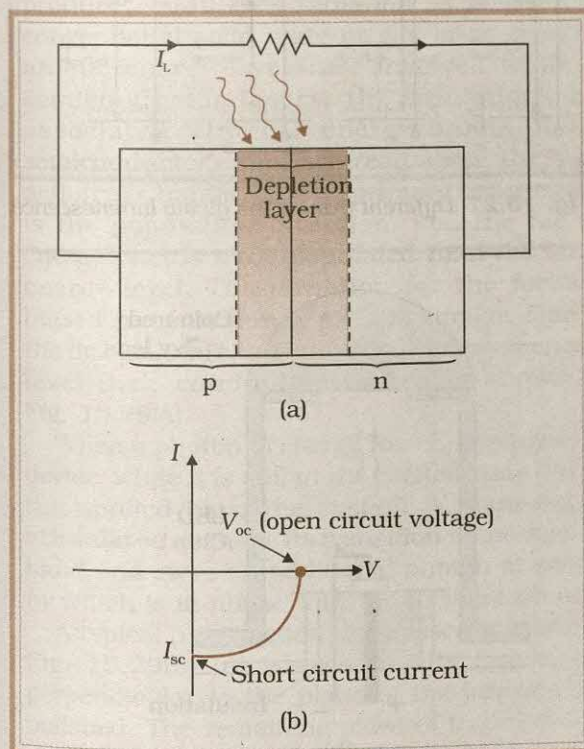


Fig. 15.25 (a) A typical illuminated p-n junction
(b) V-I characteristic of a solar cell.

When light (with $h\nu > E_g$) falls at the junction, electron-hole pairs are generated which move in opposite directions due to the junction field. Photo-generated electrons and holes respectively move towards n-side and p-side. If no external load is connected, they would be collected at the two sides of the junction giving rise to a **photo-voltage**. When external load is connected as shown in Fig. 15.25(a), a photo-current I_L flows. A typical V-I characteristic of the solar cell is shown in Fig. 15.25(b). Note that we have drawn V-I characteristics in the fourth-quadrant of cartesian co-ordinate axes. This will become clear if you carefully watch the direction of flow of the electrons in the **external circuit** and across the junction.

The materials most commonly used for solar cells are silicon (Si) and gallium arsenide (GaAs). Other important materials are cadmium sulfide (CdS), cadmium telluride (CdTe), cadmium selenide (CdSe), copper indium selenide (CuInSe_2) etc. Efficiency and cost are two important criteria for the selection of materials.

Example 15.8 Why are Si and GaAs are preferred materials for solar cells?

Answer The solar radiation received by us is shown in Fig. 15.26.

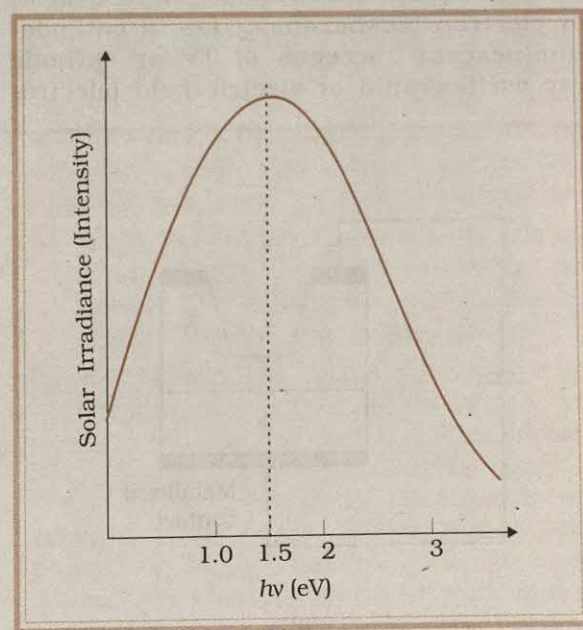


Fig. 15.26 Relative solar intensities received at different frequencies ν or energy $h\nu$.

$h\nu > E_g$. Hence, semiconductor with band gap ~ 1.5 eV or lower is likely to give better solar conversion efficiency. Silicon has $E_g \sim 1.1$ eV while for GaAs it is ~ 1.53 eV. In fact, GaAs is better (in spite of its higher band gap) than Si because of its relatively higher absorption coefficient. If we choose materials like CdS or CdSe ($E_g \sim 2.4$ eV), we can use only the high energy component of the solar energy for photo-conversion and a significant part of energy will be of no use.

The question arises: why we do not use material like PbS ($E_g \sim 0.4$ eV) which satisfy the condition $h\nu > E_g$ for ν maxima corresponding to the solar radiation spectra? If we do so, most of the solar radiation will be absorbed on the **top-layer** of solar cell and will not reach in or near the depletion region. For effective electron-hole separation, due to the junction field, we want the photo-generation to occur in the junction region only.

(c) Light Emitting Diode (LED): These are forward-biased p-n junctions which emit spontaneous radiation. You know that radiation is emitted whenever an excited electron falls from higher excited energy state to a lower energy state. For excitation, you may use thermal energy (as in incandescent lamps) or light energy (as in photo-luminescent panels in road signs) or electron bombardment (as in cathode-luminescent screens of TV or cathode ray oscillograph) or electric field (electro-

luminescence).

The possible mechanisms of **spontaneous emission** are shown in Fig. 15.27. You know that there are two distinct energy bands in a semiconductor, the conduction (higher energy) and the valence (lower energy) bands. There may also be energy bands due to donor impurities (E_D) near the conduction band or acceptor impurities (E_A) near the valence band. When electron falls from the higher to lower energy level containing holes, the energy in the form of light radiation is released. Generally, **radiative transitions** occur (sometimes non-radiative transition may also occur). The energy of radiation emitted by LED is equal to or less than the band gap of the semiconductor used as is clear from Fig. 15.27. Schematically, the construction of a LED is shown in Fig. 15.28(a). One version of the encapsulated LED (as marketed) is shown in Fig. 15.28(b).

The semiconductor used in LED is chosen

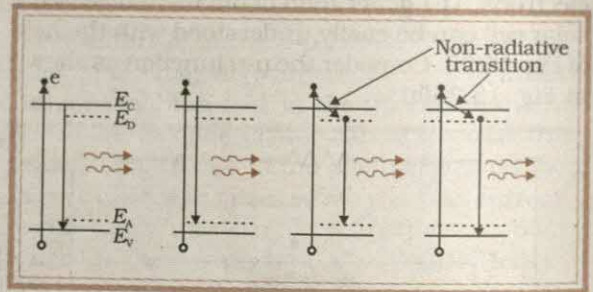
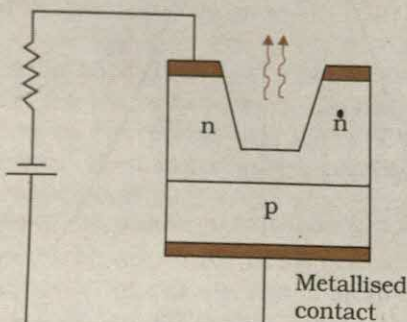
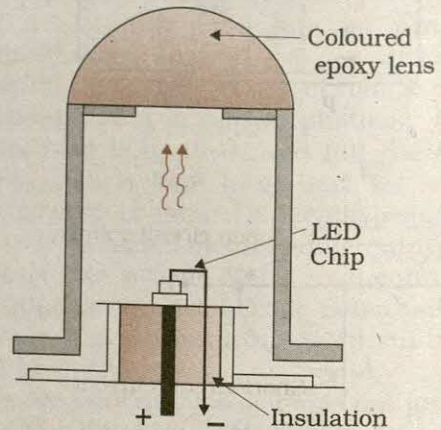


Fig. 15.27 Different transitions giving luminescence.



(a)



(b)

Fig. 15.28 (a) A typical Light Emitting Diode and (b) Encapsulated LED.

The semiconductor used in LED is chosen according to the required wavelength of emitted radiation. Visible LED's are available for red, green and orange. The visible wavelength is from $0.45\text{ }\mu\text{m}$ to $0.7\text{ }\mu\text{m}$ (Energy 2.8 eV to 1.8 eV). Therefore, the least band gap of the semiconductor for use in the visible region is 1.8 eV . Phosphorus doped GaAs ($\text{GaAs}_{1-x}\text{P}_x$) and GaP are the preferred materials. Obviously, Si ($E_g \sim 1.1\text{ eV}$) or Ge ($E_g \sim 0.7\text{ eV}$) are not suitable, since E_g is less than the minimum required E_g of $\sim 1.8\text{ eV}$.

GaAs (with $E_g \sim 1.5\text{ eV}$) alone or in conjunction with aluminium doped GaAs ($\text{Al}_x\text{Ga}_{1-x}\text{As}$) are commonly used for infrared LED's.

LED's have the following advantages over conventional incandescent lamps:

- (1) Low operational voltage and less power.
- (2) Fast action and no warm up time required.
- (3) The bandwidth of emitted light is $100\text{ }\text{\AA}$ to $500\text{ }\text{\AA}$ or in other words it is nearly (but not exactly) monochromatic.
- (4) Long life and ruggedness.

(d) Diode Laser: It is an interesting variant of LED in which its special construction helps to produce **stimulated** radiation as in laser. In conventional solid state or gas laser, discrete atomic energy levels are involved while in semiconductor lasers, the transitions are associated with the energy bands of the semiconductor. You have read about the laser action in Chapter 13. The primary requirement is the population inversion, i.e., the higher energy level is more populated than the lower energy level. The situation for the forward biased p-n junction of LED is similar. Due to the dc bias, the electrons go to the higher energy level (i.e., conduction band) as shown in Fig. 15.29(a).

When a photon of energy $h\nu = E_g$ impinges the device, while it is still in the excited state due to the applied bias, the system is immediately **stimulated** to make its transition to the valence band and gives an additional photon of energy $h\nu$ which is in phase with the incident photon.

A typical p-n junction diode laser is shown in Fig. 15.29(b). As shown in this figure, the perpendicular to the plane of the junction are polished. The remaining sides of the diode are roughened. When a forward bias is applied, a current flows. Initially at low current, there is spontaneous emission (as in LED) in all the

directions. Further, as the bias is increased, a threshold current is reached at which the **stimulated emission** occurs as explained earlier.

Due to the plane polished surfaces, the stimulated radiation in the plane perpendicular to the depletion layer builds up due to multiple reflections in the **cavity** formed by these surfaces and a highly directional coherent radiation is emitted. Diode lasers are low power lasers used as optical light source in optical communication (Chapter 16).

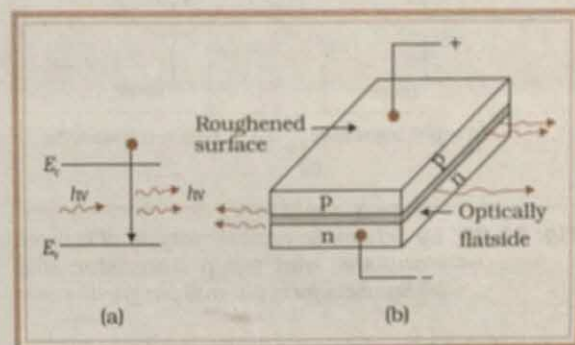


Fig. 15.29 (a) Lasing action and (b) A typical diode laser.

15.9 TRANSISTORS

15.9.1 Construction and Action

Transistor was first invented in 1948 by J. Bardeen and W.H. Brattain of Bell Telephone Laboratories, U.S.A. It consists of two p-n junctions back-to-back and is obtained by sandwiching either p-type or n-type semiconductor between a pair of opposite type of semiconductors as shown in Fig. 15.30(a). Obviously, there are two types of transistors:

- (a) **p-n-p transistor:** Here two blocks of p-type semiconductor termed as emitter and collector are separated by a thin block of n-type semiconductor (termed as base).
- (b) **n-p-n transistor:** Here two blocks of n-type semiconductor (termed as emitter and collector) are separated by a thin block of p-type semiconductor (termed as base).

You will note that all the three blocks of a transistor drawn in Fig. 15.30(a) are not equal. Further, for getting transistor action, the doping levels in the different blocks are kept different as under:

- **Emitter:** This is the left hand block of the transistor drawn in Fig. 15.30(a). It is of

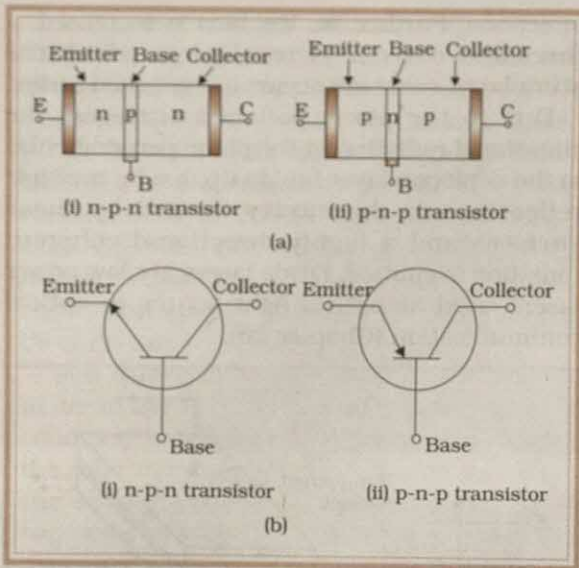


Fig. 15.30 (a) Schematic representations of an n-p-n transistor and p-n-p transistor, and (b) Symbols for n-p-n and p-n-p transistor.

moderate size and **heavily doped** semiconductor. This supplies a large number of **majority carriers** for the current flow through the transistor.

- **Base:** This is the central block. **It is very thin and lightly doped.**
- **Collector:** This collects a **major** portion of the majority carriers supplied by the emitter. The collector side is **moderately doped** and **larger** in size as compared to the **emitter**.

For understanding the action of a transistor, we have to consider the nature of depletion layers formed at the emitter-base p-n junction and base-collector p-n junction. We have to see how the charge carriers move across the transistor when proper voltages are connected at its terminals.

In general, *the emitter-base junction of a transistor is forward biased while collector base junction is reverse biased* as shown in Fig. 15.31. This figure also shows the two depletion layers formed at the emitter-base junction and collector-base junction. Since the emitter base junction is forward biased (as well as due to heavy doping of the emitter), this depletion layer will be narrow while the collector-base junction being reverse biased will be relatively wider. The forward bias voltage V_{EB} is small (0.5 to 1 V) while the reverse bias voltage V_{CB} is considerably high (5 to 15 V).

Consider the case of a biased p-n-p transistor shown in Fig. 15.31(a). As the emitter base junction is forward biased, a large number of holes (majority carriers) from p-type emitter block flow towards the base. These constitute the current through the emitter, I_E . These holes have a tendency to combine with the electrons in the n-region of the base. Only a few holes (less than 5%) are able to combine with the electrons in the base-region (giving only a very small base current, I_B) **because the base is lightly doped and very thin** (this constructional feature is the **key** of transistor

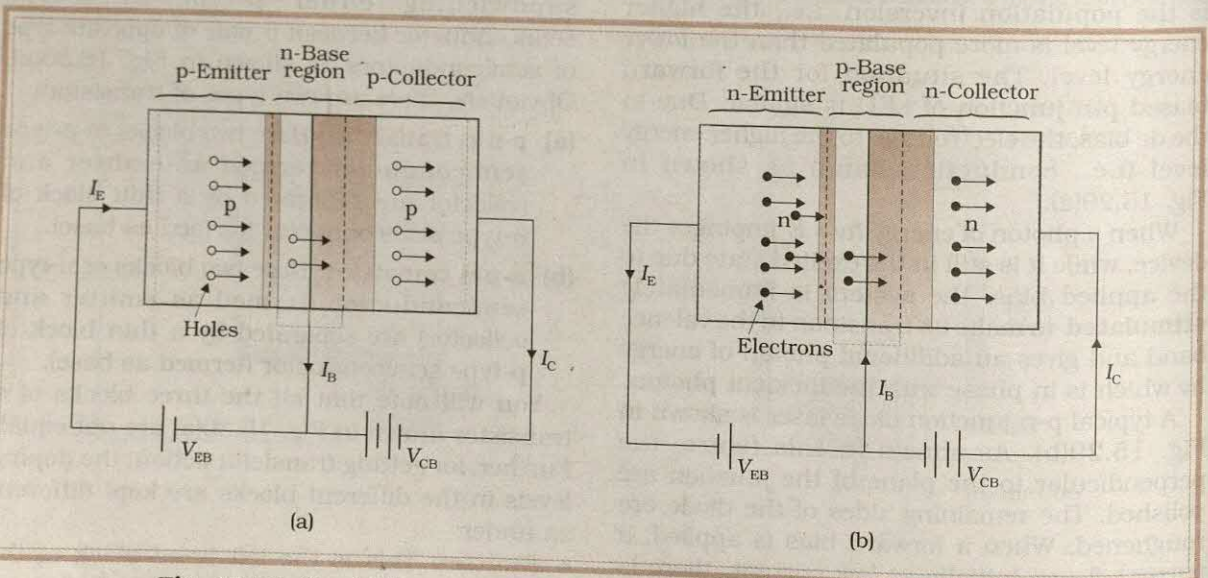


Fig. 15.31 Bias Voltage applied: (a) p-n-p transistor and (b) n-p-n transistor.

action). Most of the holes coming from the emitter are able to diffuse through the base region to the collector region. In the collector region, these holes see the favourable negative potential at the collector and hence they easily reach the collector terminal to constitute the collector current, I_C . It is obvious from the above argument (and also from a straight forward application of Kirchhoff's law to Fig. 15.31) that the emitter current is the sum of collector current and base current,

$$I_E = I_C + I_B \quad (I_C \gg I_B) \quad (15.9)$$

Similar description can be made for a biased n-p-n transistor as shown in Fig. 15.31(b). Here, the electrons (instead of holes as in p-n-p transistor) are the majority carriers supplied by the n-type emitter region which cross the thin p-base region and are able to reach the collector to give the collector current, I_C .

15.9.2 Basic Transistor Circuit Configurations and Transistor Characteristics

In any electronic circuit or device, there has to be two terminals for input and two terminals for output. In a transistor, only three terminals are available, viz., **Emitter (E)**, **Base (B)** and **Collector (C)**. Therefore, in a circuit the input/output connections have to be such that one of these (E, B or C) is common to both the input and the output. Therefore, the transistor can be **connected** in either of the three following configurations:

- Common Emitter (CE)**
- Common Base (CB)**
- Common Collector (CC)**

These configurations are shown in Fig. 15.32. In this text, we shall restrict ourselves only to CE configuration [Fig. 15.32(a)]. Any variation in the voltages on the input and output sides results in a change in the input and output currents. The variation of current on the input side with input voltage (I_B versus V_{BE}) is known as *input characteristics* while the variation in the output current with output voltage (I_C versus V_{CE}) is known as *output characteristics*.

A simple circuit for drawing the input and output characteristics of an n-p-n transistor is shown in Fig. 15.33(a). The respective input and output characteristics are shown in Fig. 15.33(b).

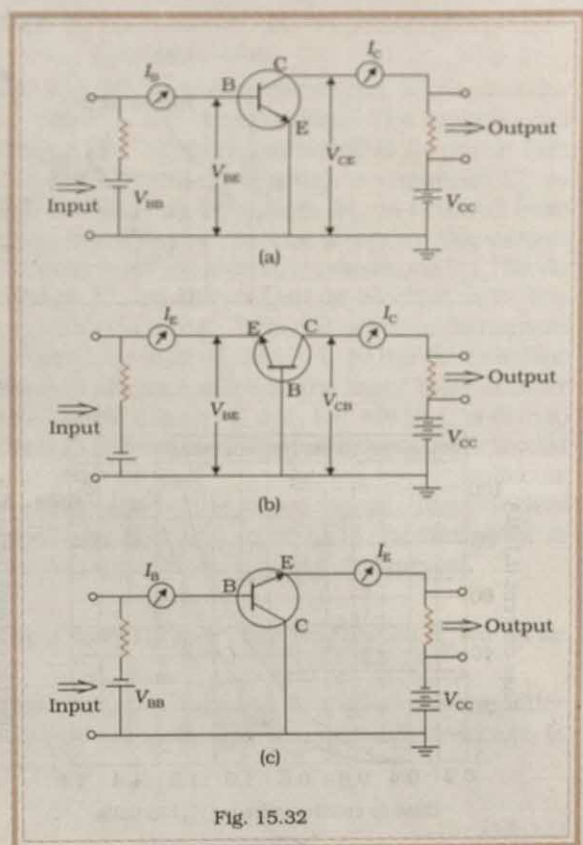


Fig. 15.32

Fig. 15.32 Different biasing configurations of a npn transistor: (a) Common Emitter (CE); (b) Common Base (CB); and (c) Common Collector (CC).

To draw the *input characteristics*, a curve is plotted between the base current (I_B) and the base-emitter voltage V_{BE} since in CE-configuration the base-emitter side constitutes the input side. In this case, V_{CE} is kept fixed while studying I_B versus V_{BE} variation. First the collector voltage V_{CE} is adjusted with the help of a rheostat, R_1 so that measurable values of I_B and I_C are obtained. This V_{CE} is kept fixed and rheostat, R_2 is varied to obtain different values of V_{BE} . The values of I_B are noted for different values of V_{BE} (keeping V_{CE} fixed). A graph I_B versus V_{BE} is plotted to get input characteristics as shown in Fig. 15.33(b)(i). Note that a microammeter is needed to read I_B since base currents are small while a milliammeter is used for reading the collector current.

To draw the *output characteristics*, the value of I_B is kept fixed (with the help of V_{BE}). For fixed value of base current on the input side (i.e., I_B), we change the value of V_{CE} and note the values

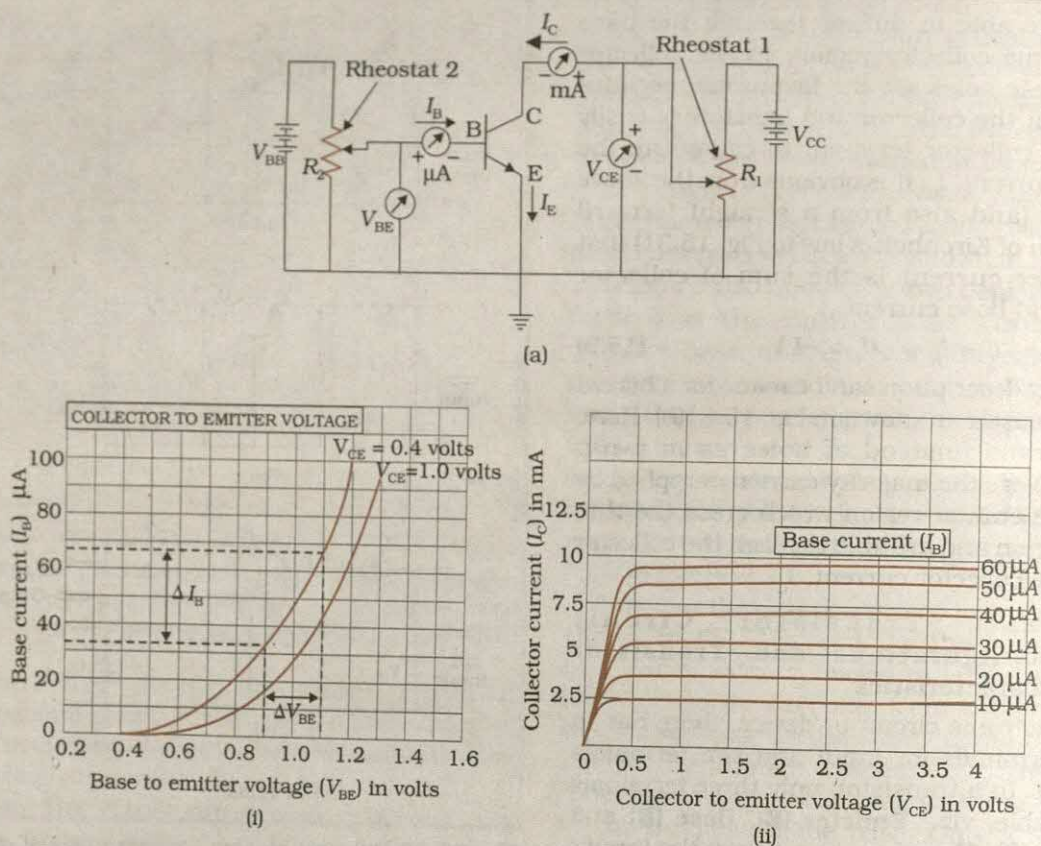


Fig. 15.33 (a) Circuit arrangement for studying the input and output characteristics of a n-p-n transistor; (b) (i) Typical input characteristics and (ii) Typical output characteristics.

of I_C . The plot I_C versus V_{CE} for different fixed values of I_B gives the output characteristics as shown in Fig. 15.33(b)(ii).

These characteristics are used to calculate the important transistor parameters as follows:

- (i) **Input resistance (r_i):** This is defined as the ratio of change in base-emitter voltage (ΔV_{BE}) to the resulting change in base current (ΔI_B) at constant collector-emitter voltage (V_{CE}).

$$r_i = \left(\frac{\Delta V_{BE}}{\Delta I_B} \right)_{V_{CE}} \quad (15.10)$$

The value of r_i is of the order of a few hundred ohms.

- (ii) **Output resistance (r_o):** This is defined as the ratio of change in collector-emitter voltage (ΔV_{CE}) to the change in collector

current (ΔI_C) at constant base current I_B .

$$r_o = \left(\frac{\Delta V_{CE}}{\Delta I_C} \right)_{I_B} \quad (15.11)$$

Since I_C changes very little with V_{CE} (after initial rise in I_C), the values of r_o are very high (of the order of 50 to 100 k Ω).

- (iii) **Current amplification factor (β):** This is defined as the ratio of the change in collector current (output current) to the change in base current.

$$\beta = \left(\frac{\Delta I_C}{\Delta I_B} \right)_{V_{CE}} \quad (15.12)$$

This is also known as **current gain**. We normally work in the region in which I_C is almost independent of V_{CE} (or varies very slowly with V_{CE}). This is called the **active region**.

Example 15.9 From the output characteristics shown in Fig. 15.33(b)(ii), calculate the values of current amplification factor of the transistor when V_{CE} is 2 V.

Answer $\beta = \left(\frac{\Delta I_C}{\Delta I_B} \right)_{V_{CE}}$

Consider any characteristics for any two values of I_B (say, 10 and 60 μA). Then for $V_{CE} = 2\text{ V}$ from the graph we have;

$$\Delta I_B = (60 - 10) \mu\text{A} = 50 \mu\text{A}$$

$$\Delta I_C \approx (9.5 - 2.5) \text{ mA} = 7.0 \text{ mA}$$

Therefore $\beta = \left(\frac{7 \times 10^{-3} \text{ A}}{50 \times 10^{-6} \text{ A}} \right) = 140.$

Example 15.10 Calculate the input resistance of the transistor operating at $V_{CE} = 4\text{ V}$ in CE configuration having its input characteristics as shown in Fig. 15.33(b)(i).

Answer From the Fig. 15.33(b)(i) and considering the dotted lines shown therein, we get

$$\Delta V_{BE} \approx (1.1 - 0.9) \text{ V} \approx 0.2 \text{ V}$$

$$\Delta I_B \approx (68 - 34) \mu\text{A} = 34 \mu\text{A}$$

$$r_i = \left(\frac{\Delta V_{BE}}{\Delta I_B} \right)_{V_{CE}} \approx 6000 \Omega$$

15.9.3 Transistor as an Amplifier (CE-Configuration)

The Fig. 15.34 shows the circuit of a transistor amplifier in CE-configuration. The input signal voltage is v_i^* . This is connected to the input side (between B and E) through a capacitor C_1 so that the biasing dc voltage V_{BB} is blocked from going towards the source of signal. The output is taken from the collector resistance R_C . The dc voltage V_{CC} in the output is blocked with the help of capacitor C_2 . Without signal, a dc current I_B flows through R_B while I_C is the dc collector current. If v_i is applied to the input base-emitter side, it will change I_B to $I_B + i_b$ where i_b is due to the signal voltage v_i . The collector current would also change to $I_C + i_c$ where i_c is the collector current due to the input signal. The effective input signal (due to v_i or i_b) to the transistor is the voltage across R_B (input resistance),

$$v_i = i_b R_B$$

The output signal voltage (v_o) across R_C would be

$$v_o = i_c R_C$$

Therefore, the voltage gain (A_v) of the amplifier (output signal divided by input signal voltage) is given by

$$A_v = \left(\frac{v_o}{v_i} \right) = \left(\frac{i_c R_C}{i_b R_B} \right) \quad (15.13)$$

Note that i_c and i_b are the respective collector and base currents due to the signal input voltage. Hence, using the definition of current amplification factor β , we can say

$$\beta = i_c / i_b$$

or $A_v = \beta (R_C / R_B) \quad (15.14)$

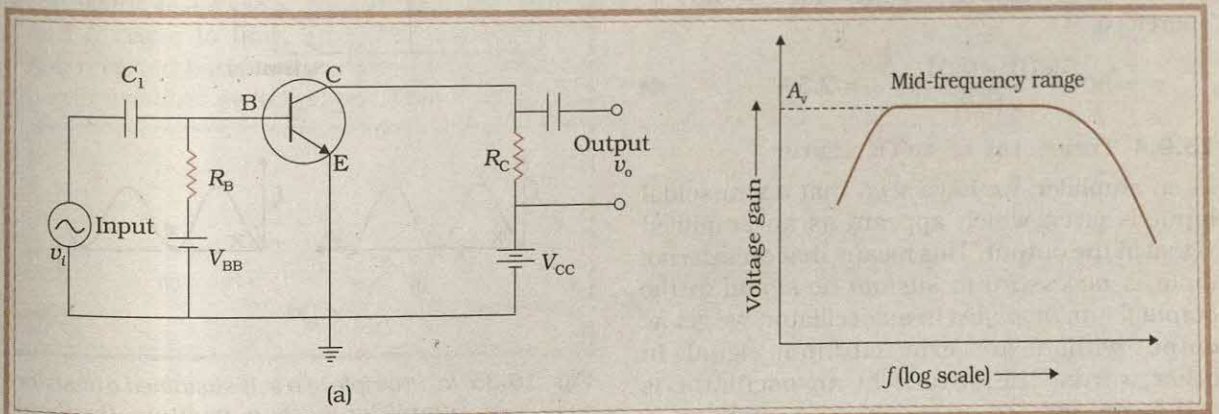


Fig. 15.34 (a) A simple circuit of a CE-transistor amplifier and (b) Frequency response curve; figure not to the scale (voltage gain versus frequency, f) of a CE amplifier.

* In electronics, generally time varying voltages and currents are denoted by lower case letters while the de biases and currents are denoted by upper case (capital) letters.

Thus, the voltage gain is related to the current amplification factor of the transistor, externally connected collector and base resistances. The above expression is only approximate since in a detailed derivation we must consider the effect of transistor parameters like base-emitter resistance, base-collector resistance, junction capacitances etc.

Because of the presence of junction and external capacitances, the voltage gain of a transistor depends upon frequency as shown in Fig. 15.34(b). This is known as **Frequency Response Curve**. You can see that gain decreases both at low frequencies as well as at high frequencies. At middle or mid frequencies, the gain remains constant. Many compensating circuits are used to get uniform gain in a broad frequency range. A detailed discussion of these is beyond the scope of this book.

Example 15.11 A transistor has a current amplification factor (current gain) of 50. In a CE-amplifier circuit, the collector resistance is chosen as 5 k Ω and the input resistance is 1 k Ω . Calculate the output voltage if input voltage is 0.01 V.

Answer From Eqs. (15.13) and (15.14),

$$A_v = \left(\frac{v_o}{v_i} \right) = \beta \frac{R_C}{R_B}$$

This,

$$v_o = \beta \frac{R_C v_i}{R_B}$$

$\beta = 50$, $R_C = 5 \text{ k}\Omega$, $R_B = 1 \text{ k}\Omega$, and $v_i = 0.01 \text{ V}$. Therefore,

$$v_o = 50 \frac{5 \times 10^3 \times 0.01}{1 \times 10^3} \text{ V} = 2.5 \text{ V}$$

15.9.4 Transistor as an Oscillator

In an amplifier, we have seen that a sinusoidal input is given which appears as an amplified signal in the output. This means that an *external input is necessary to sustain ac signal in the output for an amplifier*. In an oscillator, we get ac output without any external input signal. In other words, the output in an oscillator is **self-sustained**. To attain this, an amplifier is taken. A portion of the output power is returned back (feedback) to the input **in phase** with the starting power (this process is termed as **positive feedback**) as shown in Fig. 15.35(a).

The feedback can be achieved by inductive coupling (through mutual inductance) or LC or RC networks. Different types of oscillators essentially use different methods of coupling the output to the input (feedback network) apart from the resonant circuit for obtaining oscillation at a particular frequency. For understanding the oscillator action, we consider the circuit shown in Fig. 15.35(b) in which the feedback is accomplished by **inductive coupling** from one coil winding (T_1) to another coil winding (T_2). Note that the coils T_2 and T_1 are wound on the same core and hence are inductively coupled through their mutual inductance. As in an amplifier, the base-emitter junction is forward biased while the base-collector junction is reverse biased. Detailed biasing circuits actually used have been omitted for simplicity.

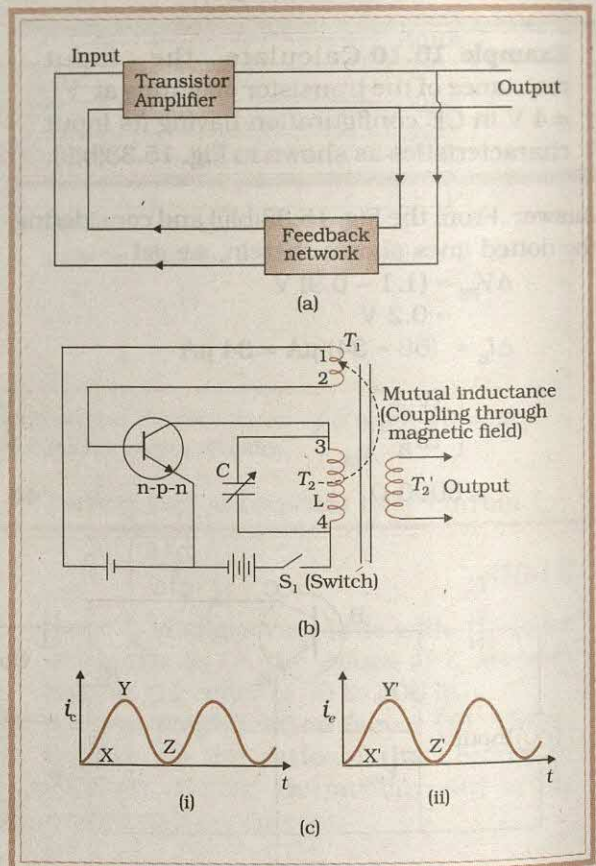


Fig. 15.35 (a) Principle of a self-sustained transistor-amplifier working as an oscillator (see text for explanation); (b) A simple LC oscillator (collector tuned); and (c) Rise and fall (or built up) of current i_c and i_e due to inductive coupling.

Let us try to understand how oscillations are built. Suppose switch S_1 is put **on** to apply proper bias for the first time. Obviously, a **surge** of collector current flows in the transistor. This current flows through the coil T_2 where terminals are numbered 3 and 4 [Fig. 15.35(b)]. This current does not reach full amplitude instantaneously but increases from X to Y , as shown in Fig. 15.35(c)(i). The inductive coupling between coil T_2 and coil T_1 now causes a current to flow in the emitter circuit (note that this actually is the 'feedback' from input to output). As a result of this positive feedback, this current (in T_1 ; emitter current) also increases from X' to Y' [Fig. 15.35(c)(ii)]. The current in T_2 (collector current) connected in the collector circuit acquires the value Y when the transistor becomes **saturated**. This means that maximum collector current is flowing and can increase no further. Since there is no further change in collector current, the magnetic field around T_2 ceases to grow. As soon as the field becomes static, there will be no further feedback from T_2 to T_1 . Without continued feedback, the emitter current begins to fall. Consequently, collector current decreases from Y towards Z [Fig. 15.35(c)(i)]. However, a decrease of collector current causes the magnetic field to decay around the coil T_2 . Thus, T_1 is now seeing a decaying field in T_2 (opposite from what it saw when the field was growing at the initial **start** operation). This causes a further decrease in the emitter current till it reaches Z' when the transistor is **cut-off**. This means that both I_E and I_C cease to flow. Therefore, the transistor has reverted back to its original state (when the power was first switched on). The whole process

now repeats itself. That is, the transistor is driven to saturation, then to cut-off, and then back to saturation. The time for change from saturation to cut-off and back is determined by the constants of the tank circuit or tuned circuit (inductance L of coil T_2 and C connected in parallel to it). The resonance frequency (f) of this tuned circuit determines the frequency at which the oscillator will oscillate.

$$f = \left(\frac{1}{2\pi\sqrt{LC}} \right) \quad (15.15)$$

In the circuit of Fig. 15.35(b), the tank or tuned circuit is connected in the collector side. Hence, it is known as **tuned collector oscillator**. If the tuned circuit is on the base side, it will be known as **tuned base oscillator**. There are many other types of tank circuits (say RC) or feedback circuits giving different types of oscillators like Colpitt's oscillator, Hartley oscillator, RC -oscillator etc., the description of which is beyond the scope of this book.

15.10 DIGITAL ELECTRONICS AND LOGIC GATES

Consider the voltage waveforms shown in Fig. 15.36.

In the waveform of Fig. 15.36(a), a continuous range of values of voltages are possible. These are **analog signals**. The electronic circuits like amplifier, oscillator etc., introduced to you in earlier sections are grouped together as **analog circuits**. In Fig. 15.36(b), you see a **pulse waveform** in which only discrete values of voltages are possible. The high level is termed as '1' while the low level is called '0'. This is closely related to the **binary system of digits** which you have already learnt. You know that in

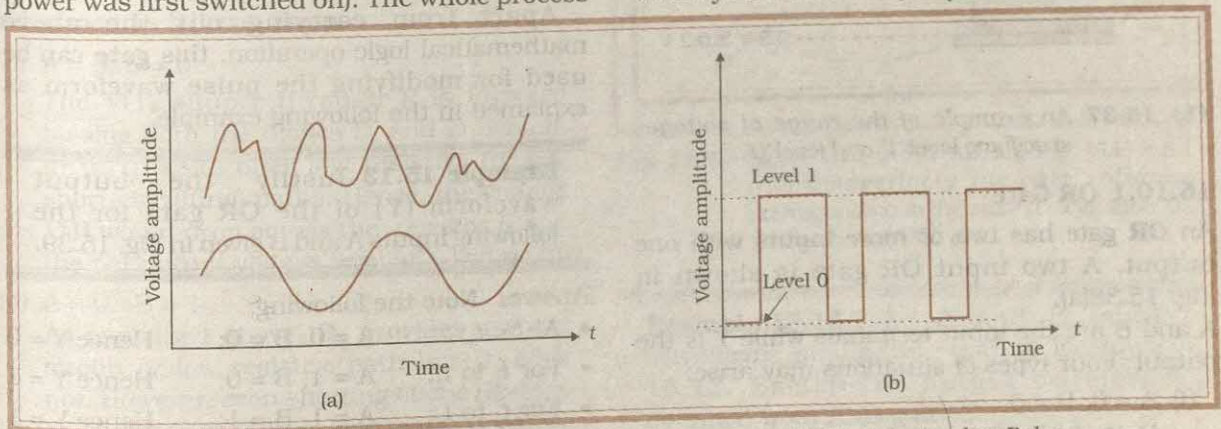


Fig. 15.36 (a) Two forms of continuous analog signals. (b) Digital signal or Pulse.

the decimal system, there are ten digits, viz. 0, 1, 2, 3, 4, 5, 6, 7, 8, and 9. In the binary system, there are only two digits 0 and 1. Using the two levels of a signal like that in Fig. 15.36(b) to represent the binary digits 0 and 1 (called **bits**), we can develop a new subject termed as **DIGITAL ELECTRONICS**.

This section is intended to provide the first step in our understanding of digital electronics. We will restrict our study to some basic building blocks of digital electronics (called **Logic Gates**) which process the 0 and 1 level signals in a specific manner.

Example 15.12 Do 0 and 1 in digital electronics designate voltages equal to 0 V and 1 V, respectively? If not, then what are the commonly used voltage levels?

Answer No! Remember that 0 and 1 levels are not the same as 0 and 1 V. In practice, the bit 0 or 1 is recognised by the presence or absence of the pulse (i.e., either at **high** or at **low** levels). Further, the digital **high** and **low** levels are specified in certain voltage levels as shown in Fig. 15.37. To avoid confusion, the voltage level of 1 and 0 are widely separated (high level is at 4 ± 1 V, low level is at 0.2 ± 0.2 V). For some devices in use, the high level may be higher than 4 V as well (like 12 V).

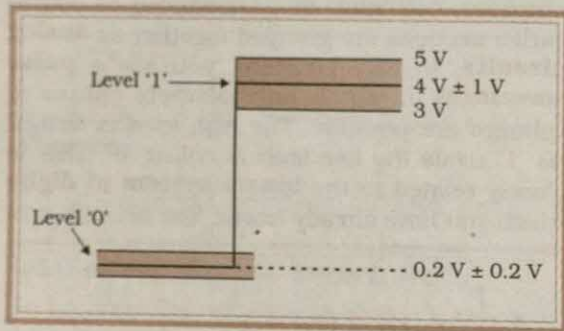


Fig. 15.37 An example of the range of voltages specifying level '1' and level '0'.

15.10.1 OR Gate

An **OR** gate has two or more inputs with one output. A two input OR gate is shown in Fig. 15.38(a).

A and B are the input terminals while Y is the output. Four types of situations may arise:

(i) $A = 0, B = 0$

Both diodes D_1 and D_2 do not conduct and

hence there will be no output voltage, i.e., $Y = 0$.

(ii) $A = 0, B = 1$

Diode D_1 does not conduct but diode D_2 conducts making the voltage level at Y as 1.

(iii) $A = 1, B = 0$

Diode D_1 conducts making output Y level as 1 even though the diode D_2 is not conducting.

(iv) $A = 1, B = 1$

Both the diodes are conducting and the level of Y is 1.

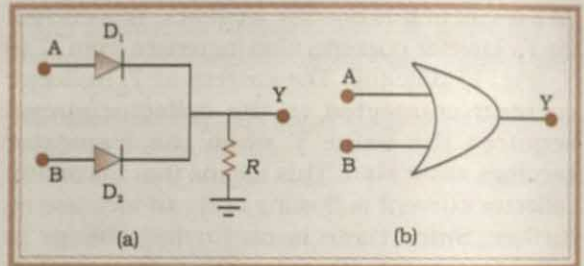


Fig. 15.38 (a) Two input OR gate. (b) Symbol for OR gate.

The above working can be stated as: **The output Y will be 1 when the input A or B or both are 1.**

This is the functional statement for OR gate. This statement can also be given in the form of a table known as **Truth Table** which is given below:

Truth Table of two input OR gate

A	B	Y
0	0	0
0	1	1
1	0	1
1	1	1

Apart from carrying out the above mathematical logic operation, this **gate** can be used for modifying the pulse waveform as explained in the following example.

Example 15.13 Justify the output waveform (Y) of the OR gate for the following inputs A and B given in Fig. 15.39.

Answer Note the following:

- At $t < t_1$; $A = 0, B = 0$; Hence $Y = 0$
- For t_1 to t_2 ; $A = 1, B = 0$; Hence $Y = 1$
- For t_2 to t_3 ; $A = 1, B = 1$; Hence $Y = 1$
- For t_3 to t_4 ; $A = 0, B = 1$; Hence $Y = 1$

- For t_1 to t_2 : $A = 0, B = 0$; Hence $Y = 0$
- For t_2 to t_3 : $A = 1, B = 0$; Hence $Y = 1$
- For t_3 to t_4 : $A = 0, B = 1$; Hence $Y = 1$

Therefore the waveform Y will be as shown in the Fig. 15.39.

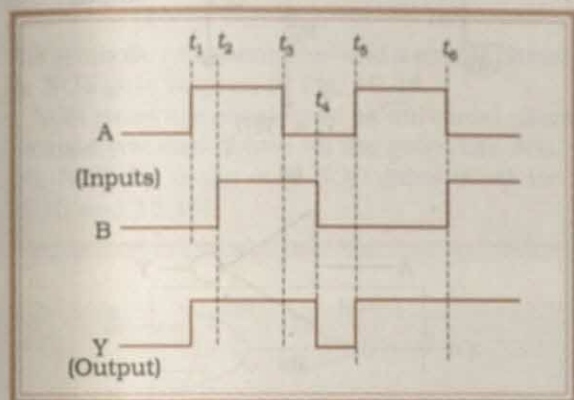


Fig. 15.39 Input and output waveforms.

15.10.2 AND Gate

An **AND** gate has two or more inputs and one output. The output Y of AND gate is 1 if all the inputs simultaneously have the state 1. This means that in the AND circuit of Fig. 15.40(a) if

- $A = 0, B = 0$; then $Y = 0$
- $A = 0, B = 1$; then $Y = 0$
- $A = 1, B = 0$; then $Y = 0$
- $A = 1, B = 1$; then $Y = 1$

In other words, if A and B are simultaneously in state 1, the output will be in state 1. A simple circuit of AND gate (for two inputs A and B) is shown in Fig. 15.40(b). The working of AND gate given in the Fig. 15.40(a) can be easily understood as follows. Suppose,

- (i) $A = 0, B = 0$

The $V(1)$ supply through R is forward biasing both the diodes D_1 and D_2 which, in turn, shall offer low resistive (nearly short-circuiting) paths. Hence, the voltage $V(1)$ would drop across the resistor R and the net output voltage level at Y will be 0.

- (ii) $A = 0, B = 1$

As explained above, D_1 provides a short circuit or low resistive path but D_2 does not. However, even shorting of one diode is enough to provide a low resistive short circuiting path for $V(1)$ connected on the

output side. Thus, the net output voltage level Y will be 0.

- (iii) $A = 1, B = 0$

In this case, the diode D_2 provides the low resistive path and D_1 does not conduct. The net result will be the output voltage level Y equal to 0 as explained above.

- (iv) $A = 1, B = 1$

None of the diodes D_1 and D_2 would conduct and there would be no drop in the voltage $V(1)$ across R . Therefore, the voltage level of Y will be $V(1)$.

The above function can be stated in the form of the following Truth Table

Truth Table of AND gate

A	B	Y
0	0	0
0	1	0
1	0	0
1	1	1

Apart from performing the above logic function, the AND gate can also modify the input pulse waveforms. The modifications would, however, be different from that produced by OR gate.

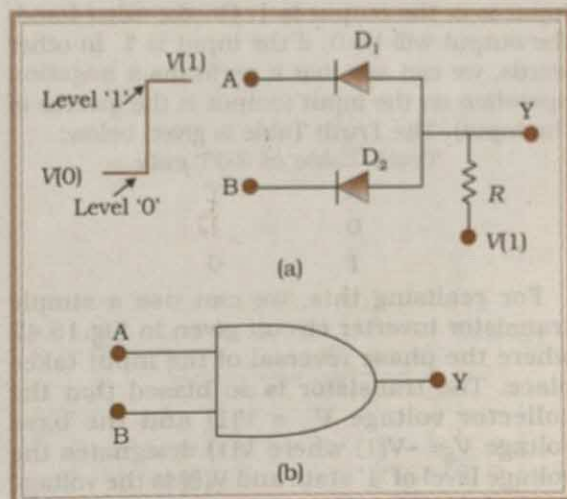


Fig. 15.40 (a) A two input AND gate. $V(1)$ and $V(0)$ respectively indicate voltages corresponding to the state '1' and state '0'. (b) Symbol of AND gate.

Example 15.14 Take A and B input waveforms similar to that in Example 15.13. Sketch the output waveform obtained from AND gate.

Answer

- For $t \leq t_1$: $A = 0, B = 0$; Hence $Y = 0$
- For t_1 to t_2 : $A = 1, B = 0$; Hence $Y = 0$
- For t_2 to t_3 : $A = 1, B = 1$; Hence $Y = 1$
- For t_3 to t_4 : $A = 0, B = 1$; Hence $Y = 0$
- For t_4 to t_5 : $A = 0, B = 0$; Hence $Y = 0$
- For t_5 to t_6 : $A = 1, B = 0$; Hence $Y = 0$
- For $t > t_6$: $A = 0, B = 1$; Hence $Y = 0$

Based on the above, the output waveform for AND gate can be drawn as given below.

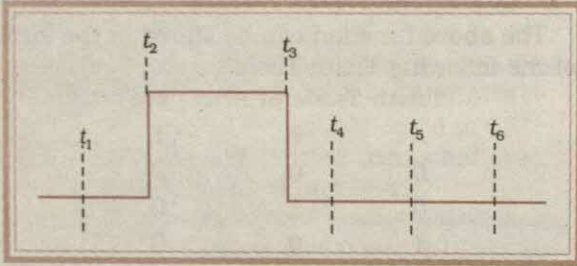


Fig. 15.41 The output waveform Y from AND gate.

15.10.3 NOT Gate

It has a single input (A) and a single output (Y). The output is **NOT** the same as the input. If the input is 0, the output is 1. On the other hand, the output will be 0, if the input is 1. In other words, we can say that it performs a **negation** operation on the input (output is the inverse of the input). The Truth Table is given below:

Truth Table of NOT gate

A	Y
0	1
1	0

For realising this, we can use a simple transistor inverter circuit given in Fig. 15.42 where the phase reversal of the input takes place. The transistor is so biased that the collector voltage $V_{cc} = V(1)$ and the base voltage $V_B = -V(1)$ where $V(1)$ designates the voltage level of '1' state and $V(0)$ is the voltage level of '0' state. The resistors (R, R_1, R_2) are so chosen that if the input is low, i.e., $V(0)$, the transistor is in the **cut-off** and hence the voltage appearing at the output will be the same as applied V_{cc} or $V(1)$. Hence, $Y = V(1)$ or state '1'. If the input is high, the transistor current is in saturation and the net voltage at the output Y is $V(0)$ or is in state 0.

The input and the output waveforms from a NOT gate are shown in Fig. 15.43.

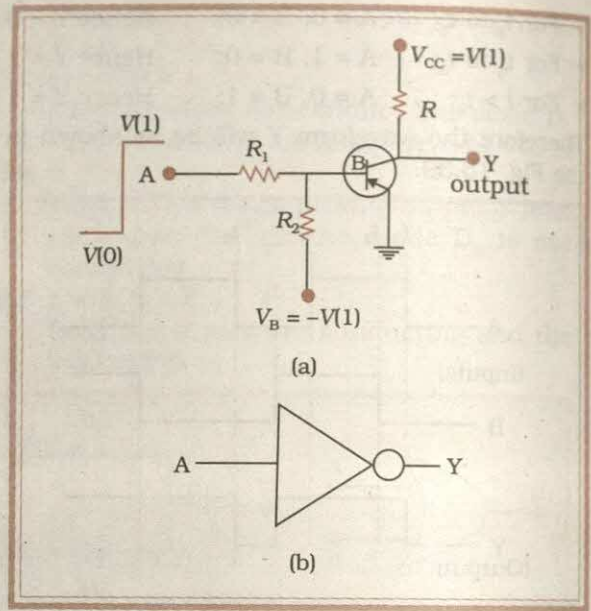


Fig. 15.42 (a) Transistor NOT gate circuit (b) Symbolic representation of NOT gate.

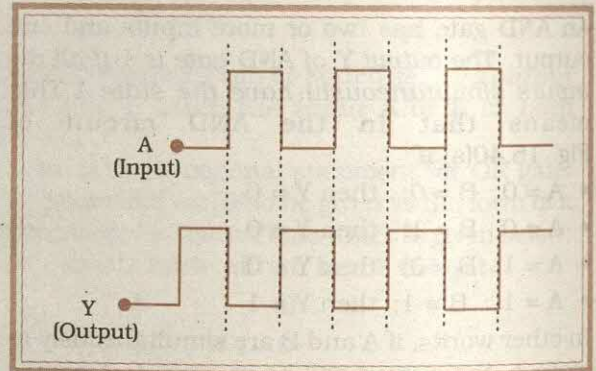


Fig. 15.43 Input and output waveforms from NOT gate.

Note that when A (input) is at high or 1 state, the output Y is at low or 0 state and vice versa.

15.10.4 NOR Gate

It has two or more inputs and one output. A negation (NOT- operation) applied **after** OR gate gives a NOT-OR gate (or simply NOR gate). This simply means that for input condition giving 1 output in OR gate will give 0 output in the NOR gate and vice-versa. For example, in a two input OR gate, if either A or B or both are 1 you get $Y = 1$ but in NOR gate it will be 0. Similarly, OR gate gives $Y = 0$ when both $A = 0$ and $B = 0$ but in NOR gate the output $Y = 1$ for such a case. Hence, Truth Table for NOR gate can be written as follows:

Truth Table for NOR gate

A	B	Y
0	0	1
0	1	0
1	0	0
1	1	0

The symbolic representation and a simple circuit for NOR gate is given in Fig. 15.44.

NOR gates are considered as **universal** gates because you can obtain all the gates like AND, OR, NOT, by using only NOR gates (Exercises 15.15 and 15.16).

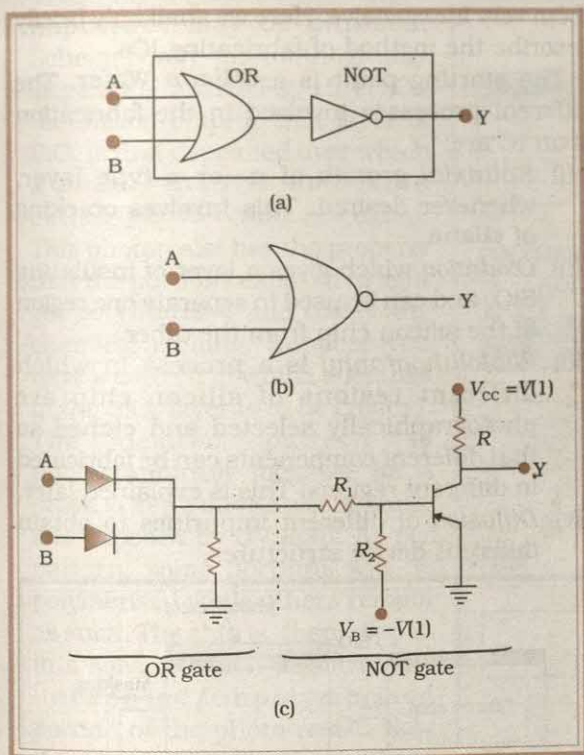


Fig. 15.44 (a) NOR gate shown as combination of OR and NOT gate. (b) Symbol for NOR gate. Note that the open circle (o) on the output side symbolises **negation**. This circle (o) distinguishes symbol of NOR from that of OR gate. (c) Complete circuit diagram of NOR gate.

15.10.5 NAND Gate

It is a combination of NOT and AND gates in which the **negation** operation (NOT) is applied after AND gate as shown in Fig. 15.45(a). The NAND is derived from NOT-AND combination. The NAND output is inverse of AND output. This simply means that for input conditions giving

1 output in AND gate will give 0 output in NAND gate and vice-versa. For example, in AND gate, the output Y is 1 only when both A and B are 1 but in NAND gate, Y will be 0 for such a case. Similarly, when both inputs are not simultaneously 1, then output is 0 in AND gate but it will be 1 in NAND gate. Hence, Truth Table for NAND gate will be as under.

Truth Table for NAND gate

A	B	Y
0	0	1
0	1	1
1	0	1
1	1	0

NAND gates are also called **Universal Gates** since by using these gates you can realise other basic gates like OR, AND and NOT (Exercises 15.18 and 15.19).

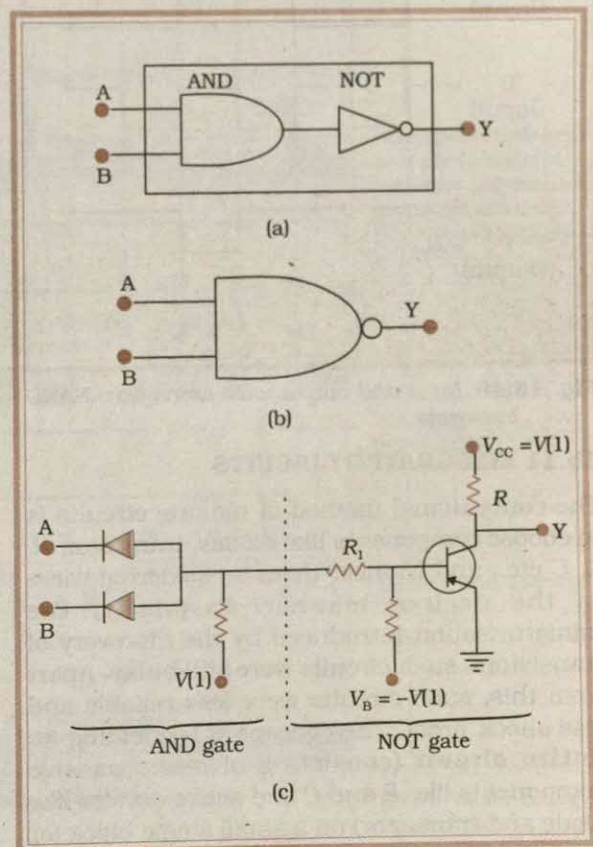


Fig. 15.45 (a) NAND gate shown as combination of NOT and AND gates. (b) Symbol of NAND gate. Note that the open circle (o) at the output side which distinguishes it from the symbol of AND gate. (c) Circuit diagram for NAND gate.

Example 15.15 Sketch the output Y from a NAND gate having inputs A and B given below:

Answer

- For $t < t_1$: $A = 1, B = 1$; Hence $Y = 0$
- For t_1 to t_2 : $A = 0, B = 0$; Hence $Y = 1$
- For t_2 to t_3 : $A = 0, B = 1$; Hence $Y = 1$
- For t_3 to t_4 : $A = 1, B = 0$; Hence $Y = 1$
- For t_4 to t_5 : $A = 1, B = 1$; Hence $Y = 0$
- For t_5 to t_6 : $A = 0, B = 0$; Hence $Y = 1$
- For $t > t_6$: $A = 0, B = 1$; Hence $Y = 1$

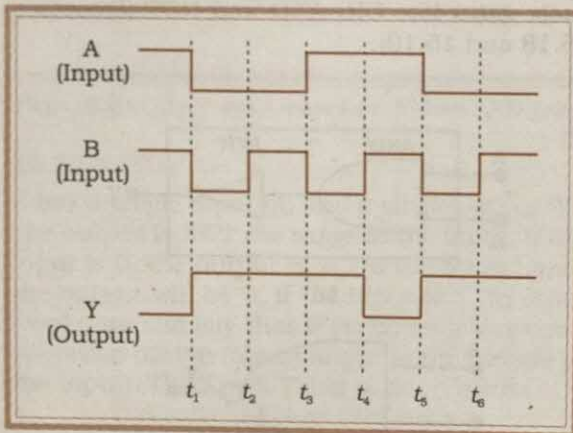


Fig. 15.46 Input and output waveforms from NAND gate.

15.11 INTEGRATED CIRCUITS

The conventional method of making circuits is to choose components like diodes, transistor, R , L , C etc., and connect them by soldering wires in the desired manner. In spite of the miniaturisation introduced by the discovery of transistors, such circuits were still bulky. Apart from this, such circuits were less reliable and less shock proof. The concept of fabricating an **entire circuit** (consisting of many passive components like R and C and active devices like diode and transistor) on a small single block (or chip) of a semiconductor has revolutionised the electronics technology. Such a circuit is known as **Integrated Circuit (IC)**. The most widely used technology is the **Monolithic Integrated Circuit**. The word **monolithic** is a combination of two greek words, **monos** means single and

lithos means stone. This, in effect, means that the entire circuit is formed on a single silicon crystal (or **chip**). The **chip** dimensions are as small as $1\text{ mm} \times 1\text{ mm}$ or it could even be smaller. Depending upon the level of integration (i.e., the number of circuit components or logic gates), the ICs are termed as **Small Scale Integration, SSI** (logic gates ≤ 10); **Medium Scale Integration, MSI** (logic gates ≤ 100); **Large Scale Integration, LSI** (logic gates ≤ 1000); and **Very Large Scale Integration, VLSI** (logic gates > 1000).

The technology of fabrication is very involved but large scale industrial production has made them very inexpensive. Here we shall only briefly describe the method of fabricating ICs.

The starting point is a **Silicon Wafer**. The different processes involved in the fabrication of an IC are;

- (i) **Epitaxial growth** of n- or p-type layer, whenever desired. This involves cracking of **silane**.
- (ii) **Oxidation** which gives a layer of insulating SiO_2 and can be used to separate one region of the silicon chip from the other.
- (iii) **Photolithography** is a process in which different regions of silicon chip are photographically selected and etched so that different components can be fabricated in different regions. This is explained later.
- (iv) **Diffusion** of different impurities to obtain different device structures.

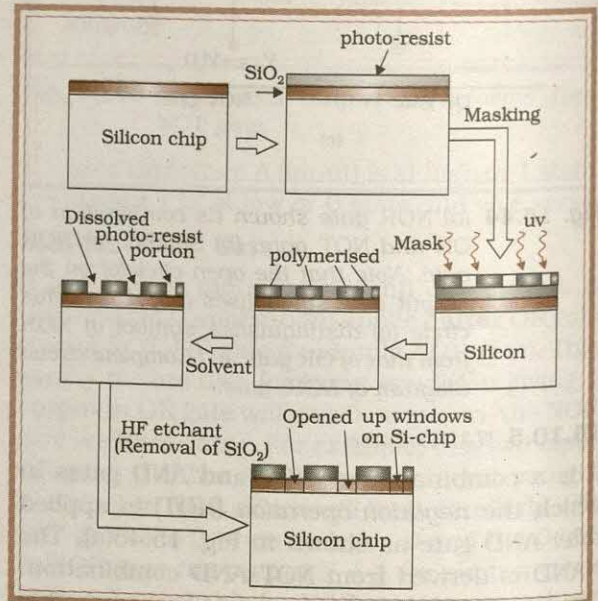


Fig. 15.47 Steps in photolithography.

(v) **Metallisation** involves deposition of metal films which inter-connect different components on a chip to obtain the circuit.

As mentioned earlier, **photolithography** is an important step in the fabrication of monolithic ICs which requires the selective removal of the SiO_2 deliberately formed over the Si-chip. This creates openings in SiO_2 layer through which impurities may be diffused. Schematically photolithography steps are shown in Fig. 15.47. On the silicon chip, a thin layer of SiO_2 is first deposited over which a **photo-sensitive emulsion** called **photo-resist** is coated. This photoresist has the property that the portions exposed to light get **polymerised** and **resist** any chemical etching. Suppose you have a pattern in which openings in the SiO_2 layer is to be created. The desired pattern is photographically focussed (or a **mask** is kept) on the photo-resist and exposed by UV light. As per pattern, some portions will be polymerised while others remain as such. The chip is, then, dipped in a solvent which dissolves the unexposed/unpolymerised portion of the photo-resist. Now it is exposed to chemical etchant which removes SiO_2 leaving exposed Si-surface beneath it through which diffusion/metallisation etc. can be carried out to get a device or circuit.

Some of the initial steps in the fabrication of ICs are summarised in Fig. 15.48. However, for getting the complete circuit (IC), the processes of masking, diffusion, oxidation, metallisation etc. have to be repeated sequentially several times.

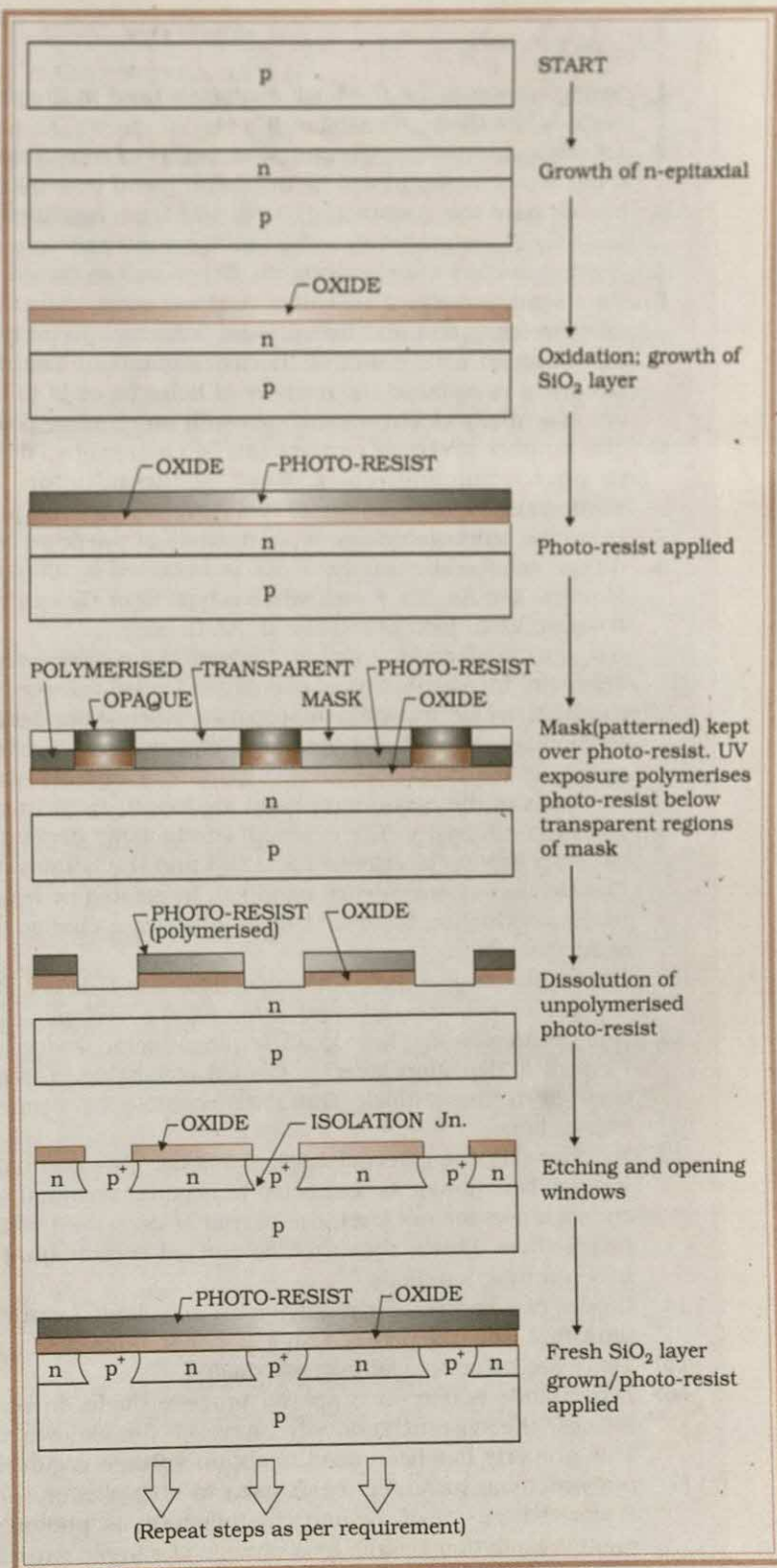


Fig. 15.48 Some preliminary steps in the fabrication of IC's.

SUMMARY

1. Semiconductors are the basic materials used in the present solid state electronic devices like diode, transistor, ICs etc.
2. Lattice structure and the atomic structure of constituent elements decide whether a particular material will be insulator, metal or semiconductor.
3. Metals have low resistivity (10^{-2} to $10^{-8} \Omega\text{m}$), insulators have very high resistivity ($>10^8 \Omega\text{m}^{-1}$), while semiconductors have intermediate values of resistivity.
4. Semiconductors are elemental (Si, Ge) as well as compound (GaAs, CdS etc.).
5. Pure semiconductors are called 'intrinsic semiconductors'. The presence of charge carriers (electrons and holes) is an 'intrinsic' property of the material and these are obtained as a result of thermal excitation. The number of electrons (n_e or simply n) is equal to the number of holes (n_h or p) in intrinsic conductors. Holes are essentially electron vacancies with an effective positive charge.
6. The number of charge carriers can be changed by 'doping' of a suitable impurity in pure semiconductors. Such semiconductors are known as extrinsic semiconductors. These are of two types (n-type and p-type).
7. In n-type semiconductors, $n_e \gg n_h$, while in p-type semiconductors $n_h \gg n_e$.
8. n-type semiconducting Si or Ge is obtained by doping with pentavalent atoms (donors) like As, Sb, P etc., while p-type Si or Ge can be obtained by doping with trivalent atom (acceptors) like B, Al, In etc.
9. $n_e n_h = n_i^2$ in all cases. Further, the material possesses an *overall charge neutrality*.
10. There are two distinct band of energies (called valence band and conduction band) in which the electrons in a material lie. Valence band energies are low as compared to conduction band energies. All energy levels in the valence band are filled while energy levels in the conduction band may be fully empty or partially filled. The electrons in the conduction band are free to move in a solid and are responsible for the conductivity. The extent of conductivity depends upon the energy gap (E_g) between the top of valence band (V_e) and the bottom of the conduction band E_c . The electrons from valence band can be excited by heat, light or electrical energy to the conduction band and thus, produce a change in the current flowing in a semiconductor.
11. For insulators $E_g > 3 \text{ eV}$, for semiconductors E_g is 0.2 eV to 3 eV, while for metals $E_g \approx 0$ (i.e., valence and conduction bands overlap).
12. p-n junction is the 'key' to all semiconductor devices. When such a junction is formed, a 'depletion layer' is formed consisting of immobile ion-cores devoid of their electrons or holes. This is responsible for a junction potential or junction energy barrier.
13. By changing the external applied voltage, junction barriers can be changed. In forward bias (n-side is connected to negative terminal of the battery and p-side is connected to the positive), the barrier is decreased while the barrier increases in reverse bias. Hence, forward bias current is more (mA) while it is very small (μA) in a p-n junction diode.
14. Diodes can be used for rectifying an ac voltage (restricting the ac voltage to one direction). With the help of a capacitor or suitable filter, a dc voltage can be obtained.
15. There are some special purpose diodes.
16. Zener diode is one such special purpose diode. In reverse bias, after a certain voltage, the current suddenly increases (breakdown voltage) in a Zener diode. This property has been used to obtain **voltage regulation**.
17. p-n junctions have also been used to obtain many photonic or opto-electronic devices where one of the participating entity is 'photon': (a) Photodiodes in which photon excitation results in a change of reverse saturation current which helps us to measure light intensity. (b) Solar cells which convert photon energy into

- 'electricity. (c) Light Emitting Diode and Diode Laser in which electron excitation by a bias voltage results in the generation of light.
18. Transistor is an n-p-n or p-n-p junction device. The central block (thin and lightly doped) is called 'Base' while the other electrodes are 'Emitter' and 'Collectors'. The emitter-base junction is forward biased while collector-base junction is reverse biased.
 19. The transistors can be connected in such a manner that either C or E or B is common to both the input and output. This gives the three configurations in which a transistor is used: Common Emitter (CE), Common Collector (CC) and Common Base (CB). The plot between I_C and V_{CE} for fixed I_B is called output characteristics while the plot between I_B and V_{BE} with fixed V_{CE} is called input characteristics. The important transistor parameters for CE-configuration are:

$$\text{input resistance, } r_i = \left(\frac{\Delta V_{BE}}{\Delta I_B} \right)_{V_{CE}}$$

$$\text{output resistance, } r_o = \left(\frac{\Delta V_{CE}}{\Delta I_C} \right)_{I_B}$$

$$\text{current amplification factor, } \beta = \left(\frac{\Delta I_C}{\Delta I_B} \right)_{V_{CE}}$$

20. Transistor can be used as an amplifier and oscillator. In fact, oscillator can also be considered as a self-sustained amplifier in which a part of output is fed-back to the input in the same phase (positive feed back). The voltage gain of a transistor

amplifier in common emitter configuration is: $A_v = \left(\frac{v_o}{v_i} \right) = \beta \frac{R_C}{R_B}$, where R_C and

R_B are respectively the resistances in collector and base sides of the circuit.

21. There are some special circuits which handle the digital data consisting of 0 and 1 levels. This forms the subject of Digital Electronics.
22. The important digital circuits performing special logic operations are called logic gates. These are: OR, AND, NOT, NAND, and NOR gates.
23. In modern day circuit, many logical gates or circuits are integrated in one single 'Chip'. These are known as Integrated circuits (IC).
24. ICs are obtained by a complex procedure involving diffusion, oxidation, photolithography, metallisation etc.

POINTS TO PONDER

1. The energy bands (E_c or E_v) in the semiconductors are space delocalised which means that these are not located in any specific place inside the solid. The energies are the overall averages. When you see a picture in which E_c or E_v are drawn as straight lines, then they should be respectively taken simply as the *bottom* of conduction band energy levels and *top* of valence band energy levels.
2. In elemental semiconductors (Si or Ge), the n-type or p-type semiconductors are obtained by introducing 'dopants' as defects. In compound semiconductors, the change in relative stoichiometric ratio can also change the type of semiconductor. For example, in ideal GaAs the ratio of Ga:As is

- 1:1 but in Ga-rich or As-rich GaAs it could respectively be $\text{Ga}_{1.1}\text{As}_{0.9}$ or $\text{Ga}_{0.9}\text{As}_{1.1}$. In general, the presence of defects control the properties of semiconductors in many ways.
3. In transistors, the base region is both narrow and lightly doped, otherwise the electrons or holes coming from the input side (say, emitter in CE-configuration) will not be able to reach the collector.
 4. We have described an oscillator as a positive feedback amplifier. For stable oscillations, the voltage feedback (V_{fb}) from the output voltage (V_o) should be such that after amplification (A) it should again become V_o . If a fraction β' is feedback, then $V_{fb} = V_o \cdot \beta'$ and after amplification its value $A(V_o \cdot \beta')$ should be equal to V_o . This means that the criteria for stable oscillations to be sustained is $A \beta' = 1$. This is known as Barkhausen's Criteria.
 5. In oscillator, the feedback is in the same phase (positive feedback). If the feedback voltage is in opposite phase (negative feedback), the gain is less than 1 and it can never work as oscillator. It will be an amplifier with reduced gain. However, the negative feedback also reduces noise and distortion in an amplifier which is an advantageous feature.

EXERCISES

- 15.1** In an n-type silicon, which of the following statement is true:
- (a) Electrons are majority carriers and trivalent atoms are the dopants.
 - (b) Electrons are minority carriers and pentavalent atoms are the dopants.
 - (c) Holes are minority carriers and pentavalent atoms are the dopants.
 - (d) Holes are majority carriers and trivalent atoms are the dopants.
- 15.2** Which of the statements given in Exercise 15.1 is true for p-type semiconductors.
- 15.3** Carbon, silicon and germanium have four valence electrons each. These are characterised by valence and conduction bands separated by energy band gap respectively equal to $(E_g)_C$, $(E_g)_{Si}$ and $(E_g)_{Ge}$. Which of the following statements is true?
- (a) $(E_g)_{Si} < (E_g)_{Ge} < (E_g)_C$
 - (b) $(E_g)_C < (E_g)_{Ge} > (E_g)_{Si}$
 - (c) $(E_g)_C > (E_g)_{Si} > (E_g)_{Ge}$
 - (d) $(E_g)_C = (E_g)_{Si} = (E_g)_{Ge}$
- 15.4** In an unbiased p-n junction, holes diffuse from the p-region to n-region because
- (a) free electrons in the n-region attract them.
 - (b) they move across the junction by the potential difference.
 - (c) hole concentration in p-region is more as compared to n-region.
 - (d) All the above
- 15.5** When a forward bias is applied to a p-n junction, it
- (a) raises the potential barrier.
 - (b) reduces the majority carrier current to zero.
 - (c) lowers the potential barrier.
 - (d) None of the above.
- 15.6** For transistor action, which of the following statements are correct:
- (a) Base, emitter and collector regions should have similar size and doping concentrations.
 - (b) The base region must be very thin and lightly doped.

- (c) The emitter junction is forward biased and collector junction is reverse biased.
- (d) Both the emitter junction as well as the collector junction are forward biased.
- 15.7** For a transistor amplifier, the voltage gain
- remains constant for all frequencies.
 - is high at high and low frequencies and constant in the middle frequency range.
 - is low at high and low frequencies and constant at mid frequencies.
 - None of the above.
- 15.8** The number of electron-hole pairs in an intrinsic semiconductor is $2 \times 10^{19} \text{ m}^{-3}$ at 27°C and E_g is 1 eV. Calculate the number of electron-hole pairs at 227°C . Given that Boltzmann constant is $8.65 \times 10^{-5} \text{ eV}$.
- 15.9** If the above semiconductor is doped by a donor impurity such that the number of conduction electrons become $2 \times 10^{24} \text{ m}^{-3}$, calculate the number of holes at 27°C . Also approximately calculate the dopant concentration.
- 15.10** In half-wave rectification, what is the output frequency if the input frequency is 50 Hz. What is the output frequency of a full-wave rectifier for the same input frequency.
- 15.11** For a CE-transistor amplifier, the audio signal voltage across the collected resistance of $2 \text{ k}\Omega$ is 2 V. Suppose the current amplification factor of the transistor is 100, find the input signal voltage and base current, if the base resistance is $1 \text{ k}\Omega$.
- 15.12** Two amplifiers are connected one after the other in series (cascaded). The first amplifier has a voltage gain of 10 and the second has a voltage gain of 20. If the input signal is 0.01 volt, calculate the output ac signal.
- 15.13** A p-n photodiode is fabricated from a semiconductor with band gap of 2.8 eV. Can it detect a wavelength of 6000 nm?

ADDITIONAL EXERCISES

- 15.14** The number of silicon atoms per m^3 is 5×10^{28} . This is doped simultaneously with 5×10^{22} atoms per m^3 of Arsenic and 5×10^{20} per m^3 atoms of Indium. Calculate the number of electrons and holes. Given that $n_i = 1.5 \times 10^{16} \text{ m}^{-3}$. Is the material n-type or p-type?
- 15.15** Write the truth table for circuit given in Fig. 15.49 below consisting of NOR gates and identify the logic operation (OR, AND, NOT) which this circuit is performing.
(Hint: $A = 0, B = 1$ then A and B inputs of second NOR gate will be 0 and hence $Y = 1$. Similarly work out the values of Y for other combinations of A and B. Compare with the truth table of OR, AND, NOT gates and find the correct one.)

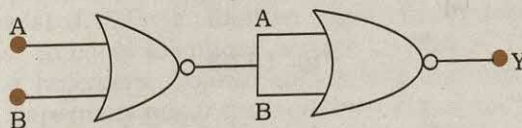


Fig. 15.49

- 15.16** Write the truth table for the circuits given in Fig. 15.50 consisting of NOR gates only. Identify the logic operations (OR, AND, NOT) performed by the two circuits.

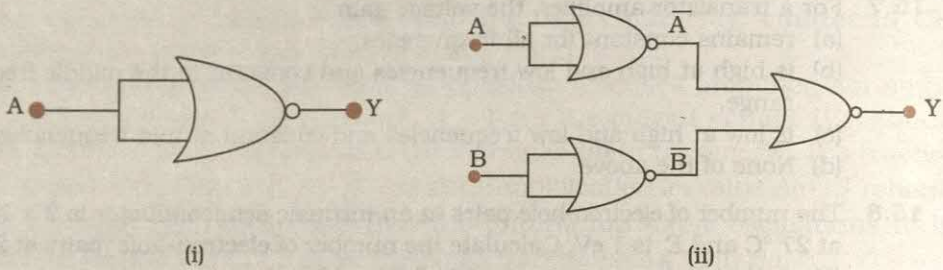


Fig. 15.50

- 15.17** You are given the two circuits as shown in Fig. 15.51. Show that circuit (a) acts as OR gate while the circuit (b) acts as AND gate.

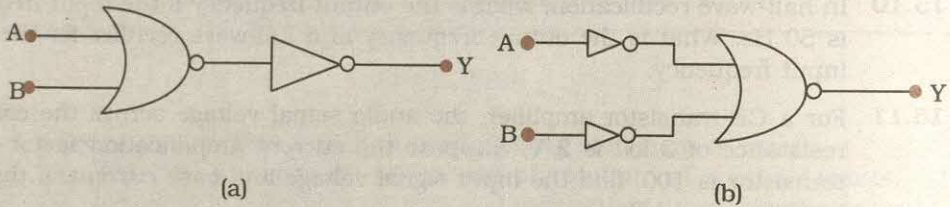


Fig. 15.51

- 15.18** Write the truth table for a NAND gate connected as given in Fig. 15.52: Hence identify the exact logic operation carried out by these circuit.

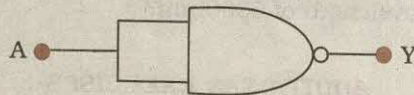


Fig. 15.52

- 15.19** You are given two circuits as shown in Fig. 15.53, which consist of NAND gates. Identify the logic operation carried out by the two circuits.

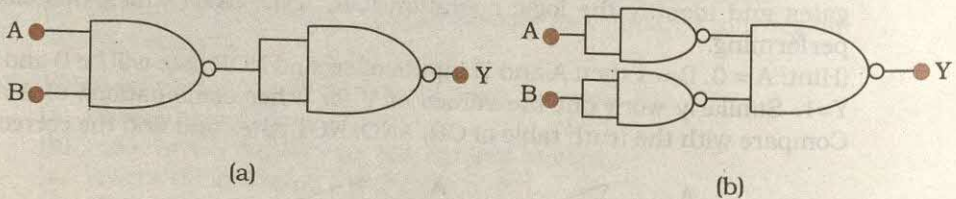
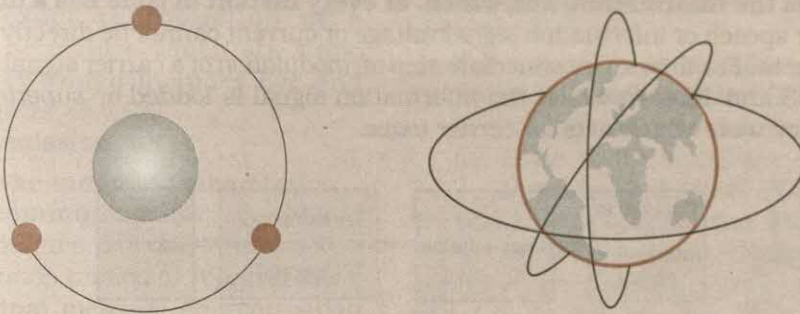


Fig. 15.53

CHAPTER SIXTEEN

COMMUNICATION SYSTEMS



16.1 INTRODUCTION

Communication of information to each other is a basic human activity. Communication refers to the faithful transfer of information or message from one point to another point where it reaches in an intelligible form. For example, one person wishes to tell something or give a message to another person sitting near him. Then he speaks and *transmits* sound waves through air medium or *channel*. On the other side, the other person *receives* the message by listening through his/her ears. This method will fail if distance between the two persons is large. In early days, long distance messages were carried by a messenger on foot/horse/cart etc. People also used coded signalling methods through smoke or flags or beating of drums. Human involvement at *every stage of communication* was necessary. However, in modern communication systems the information is first converted into electrical signals and then sent electronically. This has the advantage of speed, reliability and possibility of communicating over long distances. You are using these every day such as telephones, TV and radio transmission, satellite communication etc.

Historically, long distance communication started with the advent of telegraphy in early nineteenth century. The first trans-atlantic telegraph cable for USA-Europe communication link came up in 1850s. The milestone in trans-atlantic radio transmission in 1901 is credited to Marconi. However, the concept of radio transmission was *first* demonstrated by the famous Indian physicist J.C. Bose. Another leap in the history of communication was the development of coast to coast telephone service in USA in 1915 which used a large number of vacuum tube *repeaters*. Coaxial cables with multiple channels were laid in 1941 and by 1984 there were many major trans-atlantic channels. Satellite communication started in 1962 with the launching of **Telstar** satellite. The first *geostationary* satellite, **Early Bird**, was launched in 1965. Around 1970, high capacity optical fibre communication entered in a small way in USA, Europe and Japan. By 1988, trans-atlantic fibre-optic cables were laid.

The basic units which constitute any communication systems are shown in Fig. 16.1 together. Consider a conversation between two friends which is through *oral communication*. Can you identify the different units when you are orally communicating (talking) to your friend? Your vocal chord/tongue/lip movement *transmits* the sound waves, air is the *transmission channel* and ear is the *receiver*. In this Chapter, we would learn about **Communication Systems** where the information is in the form of electrical voltage or current signals. The key to communication system is to obtain an electrical signal voltage or current which contains the *information*. For example, a microphone can convert speech signals into electrical signals. Similarly, pressure can be sensed by **piezoelectric** sensor which gives pressure in terms of electrical signal. Light signals are converted into electrical signals by photodetectors. A device which converts a physical quantity (information) into electrical signal is known as **transducer**. It is clear that such an *electrical signal* contains the information. A signal is defined as **a single-valued function of time (that conveys the information) and which, at every instant of time has a unique value.**

Most of the speech or information signal voltage or current cannot be directly transmitted to long distances. For this, an intermediate step of *modulation* of a carrier signal is necessary (Sections 16.3 and 16.4) in which the information signal is *loaded* or *superimposed* on a high frequency wave which acts as *carrier wave*.

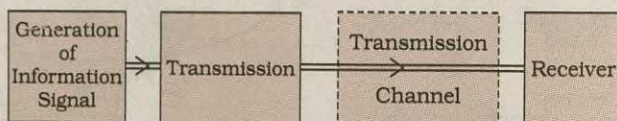


Fig. 16.1 Basic units of all communication systems.



Jagdish Chandra Bose (1858-1937)

He developed an apparatus for generating ultrashort electromagnetic waves and studied their quasi-optical properties. He was said to be the first to employ a semiconductor like galena as a self-recovering detector of electromagnetic waves. Bose published three papers in the British magazine, 'The Electrician' of 27 Dec. 1895. His invention was published in the 'Proceedings of The Royal Society' on 27 April 1899 over two years before Marconi's first wireless communication on 13 Dec. 1901. Bose also invented highly sensitive instruments for the detection of minute responses by living organisms to external stimuli and established parallelism between animal and plant tissues.

16.2 TYPES OF COMMUNICATION SYSTEMS

There is no unique way of classifying communication systems. However, for convenience, these can be classified according to the nature of information or mode of transmission or type of transmission channel used as given below:

1. According to the nature of information source

- (i) **Speech transmission** (as in radio)
- (ii) **Picture transmission** (picture including moving and live pictures)
- (iii) **Facsimile Transmission (FAX)**: This involves exact reproduction of a document or picture which are static unlike in TV which includes live and moving objects as well.
- (iv) **Data transmission.**

2. According to the mode of transmission

- (i) **Analog communication**: The signal which modulates the carrier signal for transmission is *analog* or *representative* of the original message or information to be transmitted. Note that the carrier signal may be sinusoidal or in the form of pulses. Only the **modulating signal** has to be the **analog** of the information.
- (ii) **Digital communication**: In this case, the original message or information signal is first converted into *discretised* amplitude levels and then *coded* into a corresponding sequence of binary symbols 0 and 1. Subsequently, a suitable modulation method is used.

3. According to the transmission channel

As mentioned earlier, the term transmission channel refers to the path over which the signal is being transmitted. These are:

- (i) **Line communication**
 - (a) Two wire transmission line
 - (b) Coaxial cable
 - (c) Optical fibre cable
- (ii) **Space Communication**

4. According to the type of modulation

Most of the message, information or speech signals are of low frequency which cannot be transmitted to long distances (Section 16.3). However, it is possible to radiate or transmit high frequency signals. Therefore, the information contained in the low frequency signal is somehow *loaded* or *superimposed*

on a high frequency wave which acts as *carrier* of the information. This process is known as *modulation* (Section 16.3). The transmitted carrier signal could be a **Continuous Wave (CW)** or a **Pulse**. The different schemes of modulation are explained in Section 16.3. However, for completeness these are summarised below.

For **sinusoidal continuous carrier waves**, the types of modulation are:

- (i) **Amplitude Modulation (AM)**
- (ii) **Frequency Modulation (FM)**
- (iii) **Phase Modulation**

For **pulsed carrier waves** ie various modes of modulation are:

- (i) **Pulse Amplitude Modulation (PAM)**
- (ii) **Pulse Time Modulation (PTM)**
 - (a) *Pulse Position Modulation (PPM)*
 - (b) *Pulse Width Modulation (PWM) or Pulse Duration Modulation (PDM)*
- (iii) **Pulse Code Modulation (PCM)**

PCM is the preferred modulation scheme for **digital communication** while others are more suited to analog system. We shall learn more about modulation in the next section.

16.3 MODULATION: AN IMPORTANT STEP OF COMMUNICATION SYSTEMS

What is modulation? What is the need of modulation in communication systems? How can we carry out the operation of modulation? These are some of the questions which are addressed in this section.

The purpose of a communication system is to transmit information or message signals. These are also called **baseband signals** which essentially designate the band of frequencies representing the original signal as delivered by the source of information. Note that no signal, in general, is a single frequency but it spreads over a range of frequencies called the signal **bandwidth**. Suppose we wish to transmit an electronic signal in the Audio frequency (AF) range (baseband signal frequency less than 20 kHz) over a long distance. Can we do it? The answer is 'No', because of the following problems:

1. Size of the antenna or aerial

For transmitting a signal, we need an antenna or an aerial. This antenna should have a size comparable to the wavelength of the signal (at

least $\lambda/4$ in dimension) so that the time-variation of the signal is properly sensed by the antenna. For an electromagnetic wave of frequency 20 kHz, the wavelength λ is 15 km. Obviously, such a long antenna is not possible and hence direct transmission of such baseband signals is not practical. We can obtain transmission with reasonable antenna lengths if transmission frequency is high (for example, if f is 1 MHz, and λ is 300 m). Therefore, there is a need of **translating the information contained in our original low frequency baseband signal into high or radio frequencies before transmission.**

2. Effective power radiated by an antenna

A theoretical study of radiation from a linear antenna (length l) shows that,

$$\text{Power radiated} \propto (l/\lambda)^2$$

This implies that for the same antenna length, the power radiated by short wavelength or high frequency signals would be large. Hence, the effective power radiated by long wavelength baseband signal would be small. For a good transmission, we need high powers and hence, this also points out to **the need** of using high frequency transmission.

3. Mixing up of signal from different transmitters

Another important argument against transmitting baseband signals directly is more *practical* in nature. Suppose many people are talking at the same time or many transmitters are transmitting baseband information signals simultaneously. All these signals will get **mixed** up and there is no way to distinguish between them. So, multiple user-friendly communication is not possible. This also points out towards a possible solution by using communication at high frequencies and then allotting a **band** of frequencies to each user as it is done for different radio/TV broadcast stations. This band should encompass the carrier signal frequency. For telephones, the band allotted should encompass a frequency range of 300 Hz to 3.4 kHz (general speech and audible frequency range). This in principle means that for speech signals, the bandwidth required would be 2×3.4 kHz, i.e., 6.8 kHz. Suppose you are transmitting at around 1 MHz and you are allowing 10 kHz bandwidth to each user. So, different users or transmitters can transact at frequencies: 1.00 ± 0.005 MHz,

1.01 ± 0.005 MHz, 1.02 ± 0.005 MHz, and so on; also 0.99 ± 0.005 MHz, 0.98 ± 0.005 MHz and so on. Note that symbol ± 0.005 MHz represents a bandwidth of 0.005×2 MHz or 10 kHz. Therefore, many channels get allowed if the transmission frequency is high. This argument again points the desirability of transmission at high frequencies.

The above arguments suggest that there is a **need for translating the original low frequency baseband message or information signal into high frequency wave before transmission such that the translated signal continues to possess the information contained in the original signal.** Since the high frequency signal is the signal actually transmitted and carries the information, this is why it is known as **carrier wave**. This is achieved by a process known as **modulation**. In this process, some characteristic of the transmitted carrier wave is varied in accordance with the information or message signal.

Different types of modulation depend upon the specific characteristic of the carrier wave which is being varied in accordance with the message signal. The carrier wave may be

- (i) Continuous (sinusoidal)
- (ii) Pulse

as shown in Fig. 16.2.

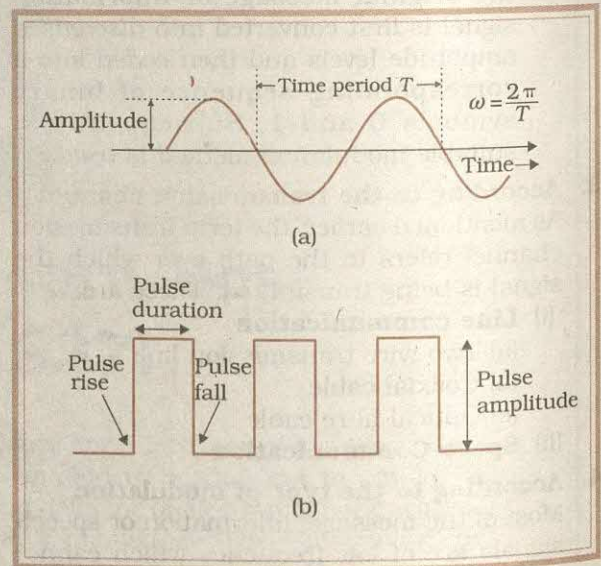


Fig. 16.2 Sinusoidal and pulse shaped signals.

You know that a sinusoidal carrier wave can be expressed as

$$E = E_c \cos (\omega_c t + \phi) \quad (16.1)$$

Obviously, its three distinct characteristics are: amplitude (E_c), angular frequency (ω_c) and phase angle (ϕ). Either of these three characteristics can be varied in accordance with the baseband signal. These result in three types of modulation: (a) Amplitude Modulation, (b) Frequency Modulation, and (c) Phase Modulation.

Similarly, the significant characteristics of a pulse are: Pulse Amplitude, Pulse Duration or Pulse Width, and Pulse Position (denoting the time of rise or fall of the pulse amplitude) as shown in Fig. 16.2. Hence, different types of pulse modulation are: (a) Pulse Amplitude Modulation (PAM), (b) Pulse Duration Modulation (PDM) or Pulse Width Modulation (PWM), and (c) Pulse Position Modulation (PPM).

16.3.1 Continuous Wave or Sinusoidal Wave Modulation

As mentioned earlier, such a modulation is of three types: Amplitude Modulation, Frequency Modulation and Phase Modulation. These three forms of modulation are shown in Fig. 16.3 (along with the carrier wave signal and modulating signal). The implications of these are described below.

(i) **Amplitude Modulation (AM):** In this mode of modulation, as shown in Fig. 16.3(c) the **amplitude** of the carrier signal varies in accordance with the **modulating signal** (message or information signal). The high frequency carrier wave is shown in Fig. 16.3(a) while the low frequency information signal (termed as modulating signal) is shown in Fig. 16.3(b). You can see that in the amplitude modulated carrier wave, the amplitude is not constant. When the amplitude (instantaneous) of modulating wave is increasing, the amplitude of the carrier wave increases and vice-versa. Thus, the *amplitude* of the modulated carrier is not constant but its *envelope* has similar sinusoidal variation as that of the low frequency or modulating signal. In simple words, the carrier modulated wave is loaded with the information contained in the low frequency message signal.

(ii) **Frequency modulation (FM):** In this mode of modulation, as shown in Fig. 16.3(d), frequency of the carrier signal varies in accordance with the modulating signal. You

can see that the amplitude of the carrier wave is fixed while its frequency is changing. When the instantaneous amplitude of the modulating voltage is large, the instantaneous carrier frequency is higher. It is lower when the modulating voltage is small.

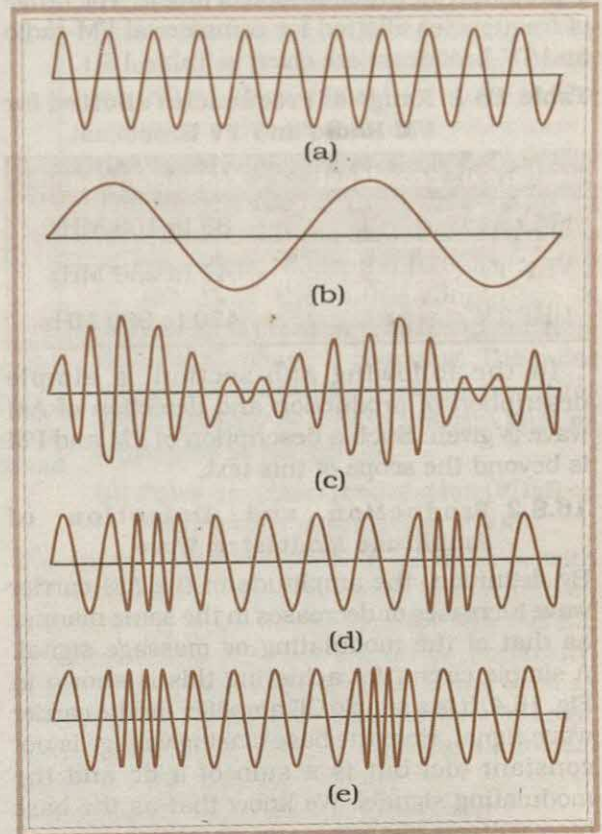


Fig. 16.3 Modulation of a carrier wave: (a) a sinusoidal carrier wave; (b) a modulating signal; (c) amplitude modulation; (d) frequency modulation; and (e) phase modulation.

(iii) **Phase modulation (PM):** Here the **phase angle** ϕ of the carrier signal varies in accordance with the modulating voltage. In the, Fig. 16.3(e), you are seeing it as a varying frequency because the phase term is an *angle* term like ω in the equation of sinusoidal wave.

Amplitude modulation is widely used for commercial broadcast of voice signals. The carrier wave frequency ranges from approximately 0.5 to 2.0 MHz. AM signals are *noisy* as atmospheric (like lightning discharge) or man-made electrical noise signals significantly affect this. Frequency modulation (FM) gives

better quality transmission and has a larger bandwidth. In FM signals, the intelligence (information or message signal) is in the form of frequency variations and, therefore, the atmospheric or man-made noises (which are generally amplitude changes) do little harm. It is preferred for transmission of music. The range of frequencies allotted for commercial FM-radio and TV-broadcast are given in Table 16.1.

Table 16.1 Range of Frequencies allotted for FM Radio and TV Broadcast

Nature of broadcast	Frequency band
FM radio	88 to 108 MHz
VHF TV	47 to 230 MHz
UHF TV	470 to 960 MHz

In the following sub-section, a simple description of production and detection of AM wave is given. Such a description of FM and PM is beyond the scope of this text.

16.3.2 Production and Detection of Amplitude Modulated Wave

By definition, the amplitude of the AM carrier wave increases or decreases in the same manner as that of the modulating or message signal. A simple circuit for achieving this is shown in Fig. 16.4. It is a simple CE amplifier for the carrier wave signal. Here the base biasing voltage is not constant (dc) but is a sum of a dc and the modulating signals. We know that as the base bias voltage changes, the amplification will change and the output voltage will be a carrier signal varying in amplitude in accordance with the biasing modulating voltage, thus giving an AM signal.

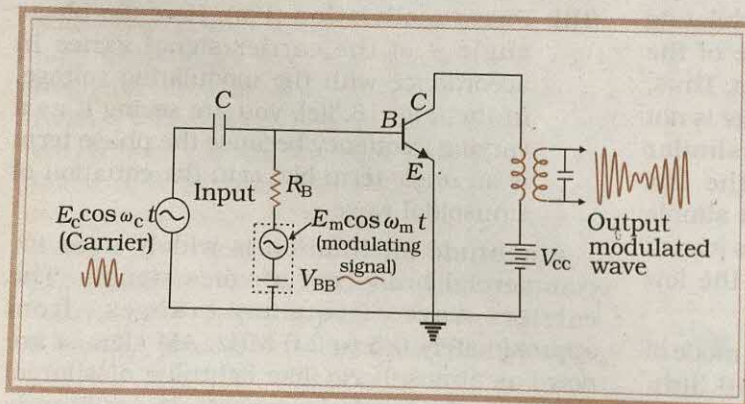


Fig. 16.4 Circuit of an amplitude modulator (Base modulator).

From the modulated signal, we can **demodulate** or **detect** the modulating or message signal by using a diode and a suitable capacitor filter as shown in Fig. 16.5(a). The AM wave input is shown in Fig. 16.5(b)(i). It appears at the output of the diode (across A'B') as a *rectified* wave (since a diode conducts only in the positive half cycle) as shown in Fig. 16.5(b)(ii). This rectified wave after passing through the RC network does not contain the radio frequency carrier component as shown in Fig. 16.5(b)(iii). Instead, it has only the envelope of the modulated wave. This can be qualitatively understood in terms of charging and discharging of the capacitor C connected in parallel to the load R. The capacitor charges till the positive peak of the rectified carrier signal achieves its maximum amplitude. When the amplitude starts decreasing, the capacitor tries to discharge its charge through R. However, this takes time. It has been found that the charge on the capacitor discharges to $1/e$ of its value (where e stands for exponential) in time.

$$t = RC \quad (16.2)$$

This value of RC is known as the **time constant**. In the actual circuit, the value of RC is chosen such that

$$\frac{1}{f_c} \ll RC \quad (16.3)$$

where f_c is the frequency of the carrier signal. Under this condition, the capacitor charges to the positive peak of the amplitude modulated wave and then starts discharging. Since the time period of the carrier frequency is smaller than the time constant RC , the voltage across RC goes down very little (enough time is not available for discharging) before the next peak of the signal starts charging the capacitor again. In this way, the output voltage across RC circuit roughly follows the envelope of the modulated waveform (which is representative of the input modulating message signal). Thus, we have achieved demodulation or detection.

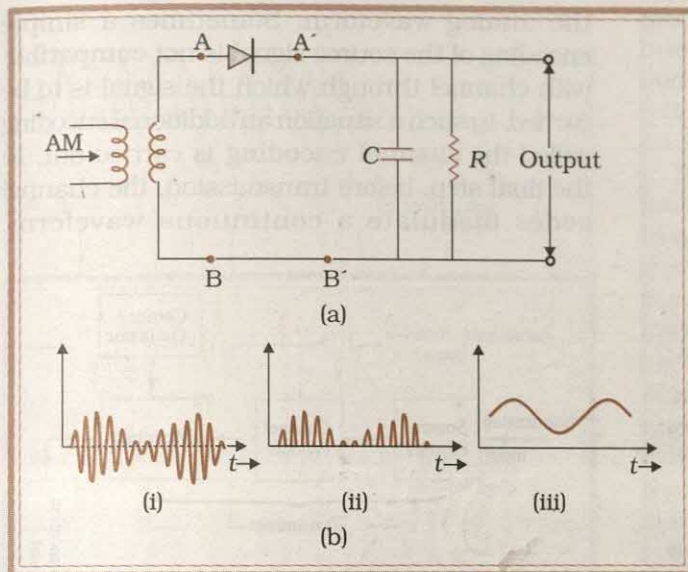


Fig. 16.5 (a) An AM demodulator and (b) Input and output wave forms in AM demodulator: (i) AM wave input; (ii) rectified wave output; and (iii) output wave without radio frequency component.

Example 16.1 In a diode AM-detector, the output circuit consists of $R = 1 \text{ k}\Omega$ and $C = 10 \text{ pF}$. A carrier signal of 100 kHz is to be detected. Is it good? If yes, then explain why? If not, what value would you suggest?

Answer For demodulation:

$$\frac{1}{f_c} \ll RC$$

$$\frac{1}{f_c} = \frac{1}{100 \times 10^3} = 10^{-5} \text{ s}$$

$$RC = 10^3 \times 10 \times 10^{-12} \text{ s} \\ = 10^{-8} \text{ s}$$

We see that $1/f_c$ here is not less than RC as required by the above condition. Hence, this is not good.

For a satisfactory circuit, try $C = 1 \text{ }\mu\text{F}$. The product RC would now be $10^3 \times 10^{-6} \text{ s} (=10^{-3} \text{ s})$. This

is much larger than $\frac{1}{f_c}$ as required for detector.

As a practical activity, you try to find out the frequency of the radio transmitter nearest to you. Following the above illustrative example, you can then calculate the suitable combination of R and

C . Try to make a circuit as shown in Fig. 16.5(a) with a simple diode and test whether you can listen to the radio with a *headphone* in the output.

16.3.3 Pulse Modulation

Here the carrier wave is in the form of *pulses*. Essentially, Pulse Modulation (PM) is an analog process since the modulating or information signal is analog. The common pulse modulation systems are:

(i) Pulse amplitude modulation (PAM):

The **amplitude** of the pulse varies in accordance with the modulating signal as shown in Figs. 16.6(c) and 16.6(d). It is easy to see that the *pulse amplitude* is increasing or decreasing as the modulating sinusoidal voltage is changing. The pulse amplitude modulation could be either single polarity or double polarity PAM as illustrated in the figures.

(ii) Pulse duration modulation (PDM) or Pulse width modulation (PWM):

The **pulse duration** varies in accordance with the modulating signal as shown in Fig. 16.6(e). Note that the pulse duration or the width of the original unmodulated signal is constant. However, you can easily see that in the PDM as shown in Fig. 16.6(e), the pulse duration is large, when amplitude of the modulating sinusoidal signal is large and vice-versa.

(iii) Pulse position modulation (PPM):

The **pulse position** (i.e., time of rise or fall of the pulse) changes with the modulating signal as shown in Fig. 16.6(f). The dashed lines in this figure show the original position of the pulse. Note that in PPM these **positions** have shifted in time. The shift is more if modulating amplitude is high and vice-versa.

Note that all types of modulations (discussed in Sections 16.3.1, 16.3.2 and 16.3.3) are *analog*; since in all these cases the *modulated signal* has a characteristic which is variable and proportional to the modulating voltage.

There is another *variant* of pulse modulation, called as **Pulse Code Modulation (PCM)**, which is **digital** in nature. Hence, the modulating or message signal is first *discretised* or *digitised* in binary code (or quantised, see Section 16.4) which, in turn, is used to produce the modulated signal for transmission.

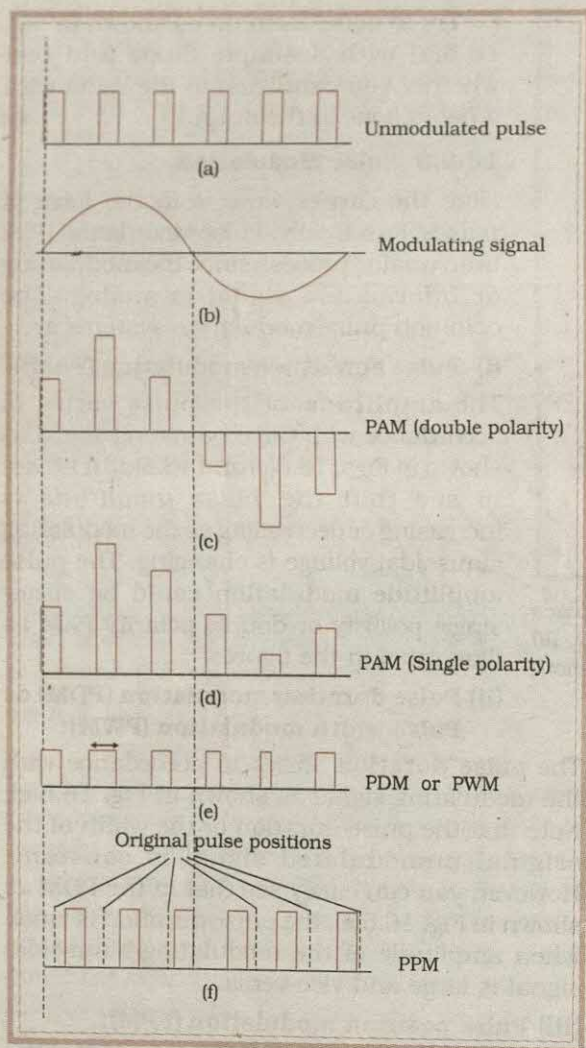


Fig. 16.6 Different types of pulse modulated signals.

16.4 DIGITAL COMMUNICATION AND QUANTISATION OF MESSAGE SIGNAL

In digital communication, the modulating signals are discrete and are *coded* representation of the message signal or information to be transmitted. However, in analog communication the signals are **continuous** signals and are essentially **analogous** to the message or information. A digital communication system vis-à-vis analog system is shown in Fig. 16.7. There are a number of *encoding steps* in the digital communication, which makes its circuitry complicated. The digital communication ensures relatively more error- and noise-free communication. The source encoder converts the information into binary code. A part of the job of the source encoder is to first *digitise*

the analog waveform. Sometimes a simple encoding of the source signal is not **compatible** with channel through which the signal is to be carried. In such a situation an additional encoding called the **channel encoding** is carried out. In the final step, before transmission, the channel codes modulate a continuous waveform.

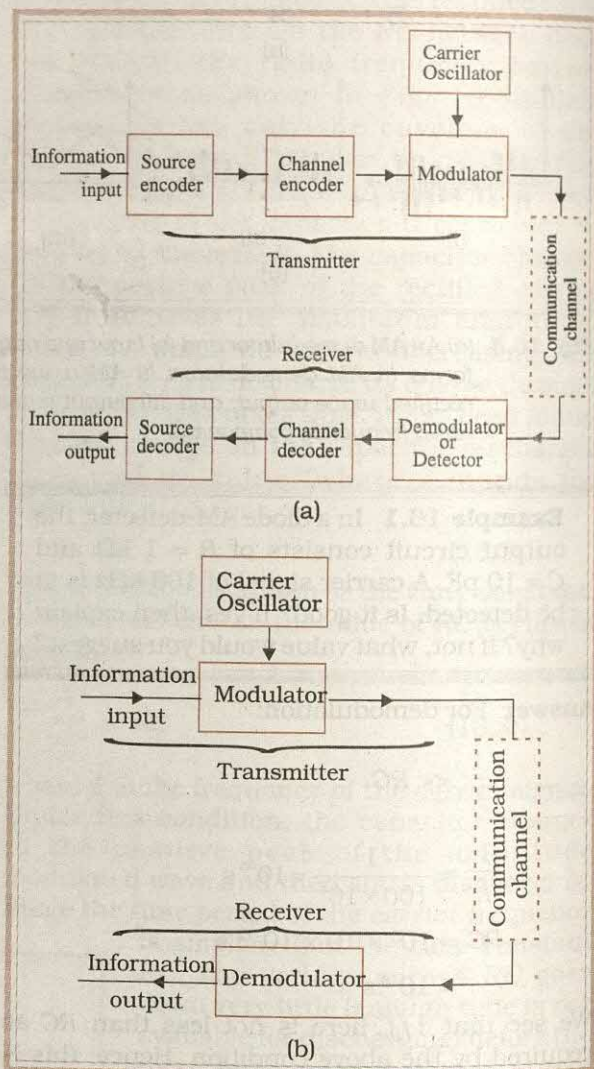


Fig. 16.7 (a) Digital and (b) Analog communication system.

The common modulation techniques employed for the **digital data** are:

- (a) **Amplitude Shift Keying (ASK)**
- (b) **Frequency Shift Keying (FSK)**
- (c) **Phase Shift Keying (PSK)**

The ASK and FSK are shown in Fig. 16.8. The carrier is a continuous radio frequency signal.

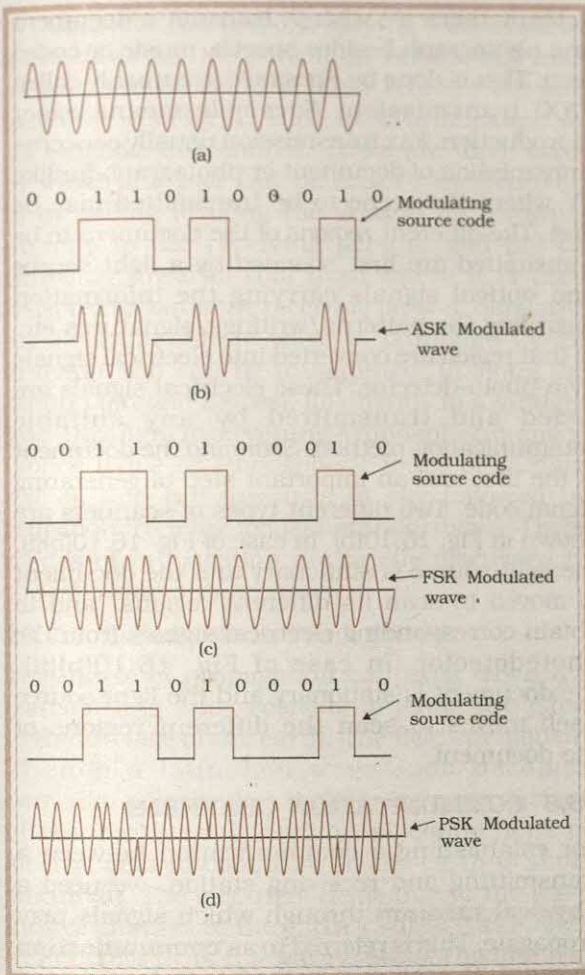


Fig. 16.8 Amplitude and Frequency Shift Keying used for digital data transmission.

The data (or message) is given in the form of a code. In ASK, the *amplitude* of the carrier changes in accordance with the data signal. In binary data, the amplitude values are 0 or 1. Hence, in ASK the carrier amplitude is either 0 or has the original value of the carrier. For 1, the original carrier of constant amplitude is present while it is absent for 0 as shown in Fig. 16.8(b). Similarly, for FSK the frequency of carrier increases for 1 while it remains unaffected for data signal 0. In PSK, it is the phase which changes with 0 and 1.

The digital data is 0 and 1. Obviously, it has only **discrete** values and can be said to be *quantised* in the sense that it is either 0 or 1. Continuous analog values are not present. Most of the information or message signals are *analog*. Then, the basic question is how an analog signal can be represented as coded pulse? The detailed

procedure of *coding* is beyond the scope of this book. However, one necessary step is the *quantisation* of the analog signal. The process of quantisation is illustrated in Fig. 16.9. Consider an analog signal $a(t)$.

This signal is divided into different steps such that the output of the quantiser has a **staircase** like shape. Note that the analog input varies smoothly while the quantised signal holds itself at one or the other of the fixed levels V_1, V_2, V_3, \dots . Thus, the signal either does not change or changes abruptly by a *quantum jump* of size S called the *step-size*. Therefore, the quantised signal is an approximation of the original signal. The quality of the approximation may be improved by reducing the size of the steps. Each quantised signal level, subsequently, can be assigned a binary code or any other code. During the quantisation step, the noise gets eliminated which makes the digital communication more reliable than the analog.

16.5 DATA AND DOCUMENT TRANSMISSION: FAX AND MODEM

The term **data** is applied to a representation of facts, concepts or instructions suitable for communication, interpretation or processing by human beings or by automatic means. Data in most cases consists of pulse-type of signals. Even an analog signal can be converted into a series of coded pulses. With the advent of computer, data transmission from one machine to another has become very important. *Modems* are used to interface two digital sources/receivers. The name *modem* is a contraction of the terms **MOD**ulator and **DEM**odulator (**MO** + **DEM**). As the name implies, both the functions are included in a single unit. Modems are placed at both ends of the communication circuit as shown in Fig. 16.10(a). The modem at the transmitting station changes the digital output from a computer (or any other business machine) to a form which can be easily sent

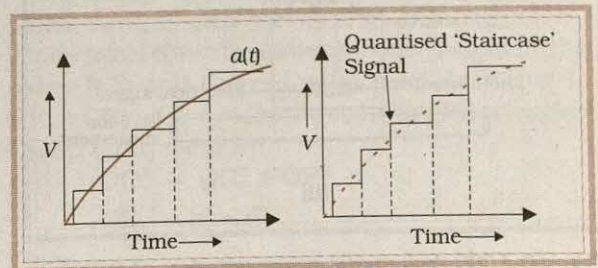


Fig. 16.9 Process of signal quantisation.

via a communication channel, while the receiving modem reverses the process. When used in the transmitting mode, the modem accepts digital data and converts it to analog signals for use in modulating a carrier signal (FSK or PSK as explained in the earlier Section 16.4). At the receiver end of the system, the carrier is demodulated to recover the data.

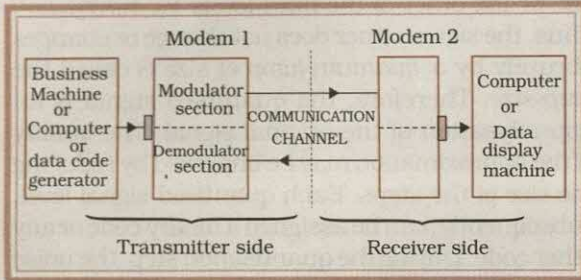


Fig. 16.10(a) Data communication circuit using MODEMS.

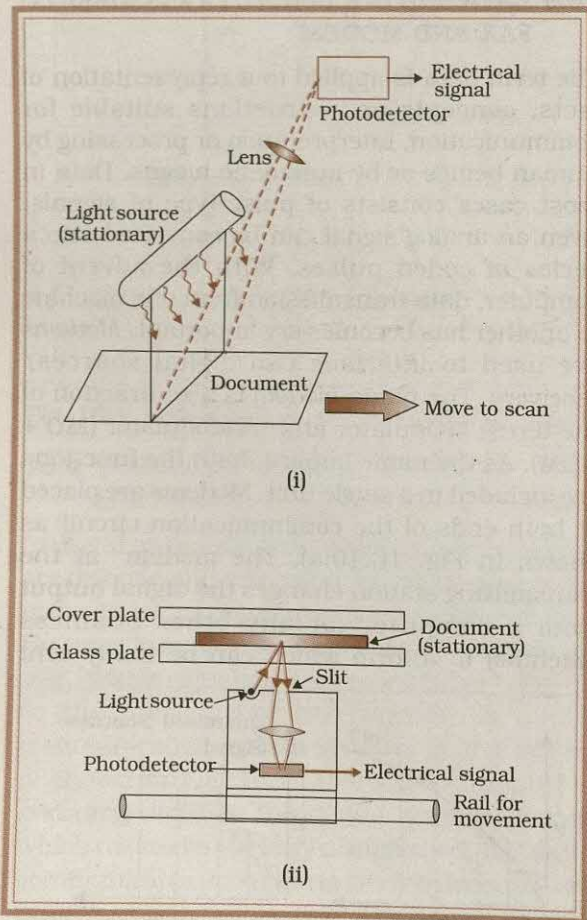


Fig. 16.10(b) Schematic representation of scanners used in Fax machine.

Many times we wish to transmit a document or a photograph besides speech, music or coded data. This is done by *Facsimile* (commonly called FAX) transmission. *Facsimile* means exact reproduction. Fax transmission usually concerns transmission of document or photograph (unlike TV where the scene to be transmitted may be live). The different regions of the document to be transmitted are first scanned by a light source and optical signals carrying the information regarding the patterns/writings/signatures etc. in that region are converted into electrical signals by a photo-detector. These electrical signals are coded and transmitted by any suitable communication method. *Scanning* the document is the first and an important step of generating signal code. Two different types of scanners are shown in Fig. 16.10(b). In case of Fig. 16.10(b)(i), the light source is stationary and the document is moved to scan its different regions and to obtain corresponding electrical signals from the photodetector. In case of Fig. 16.10(b)(ii), the document is stationary and the light source itself moves to scan the different regions of the document.

16.6 COMMUNICATION CHANNELS

For establishing a communication between a transmitting and receiving station, we need a **physical medium** through which signals may propagate. This is referred to as **communication channel**. One simple method is to launch and to guide the signals along a conducting physical path (or line) to the receiver. This mode of communication is broadly referred to as **line communication**. The other method is to transmit the signal *freely* in space by using an antenna and receive at the other end by *intercepting* the signal with the help of another antenna. This mode of communication is termed as **space communication**. Thus, principally there are two types of communications.

(i) **Space communication**

(ii) **Line communication**

A new dimension recently added to the **space communication** is **satellite communication**.

These are briefly described in the subsequent sections.

16.7 SPACE COMMUNICATION

The communication process utilising the physical space around the earth is termed as **space communication**.

Consider two persons playing with a ball in a closed room. One person throws the ball (acts as transmitter) and the other receives the ball (acts as receiver). Ball is the message or information transmitted. There are three ways in which the ball can be sent to the receiver; (i) by rolling it along the ground (ii) throwing directly to the receiving person, and (iii) by throwing towards roof from where it is then reflected towards the receiving person. Similarly, an electromagnetic wave is launched by an antenna and can be transmitted: (i) along the ground (*ground waves*), (ii) directly in a straight line through intervening tropospheric space (*space wave* or *tropospheric wave* or *surface wave*), and (iii) upwards in sky followed by reflection from the *ionosphere* (*sky wave*). These modes of propagation are briefly discussed in the following sub-sections.

16.7.1 Ground Wave Propagation

This mode of propagation can exist when the transmitting and receiving antenna are close to the surface of the earth. The field component of such a launched wave soon becomes *vertically polarised* as it glides over the surface of the earth (any horizontal component of electric field in contact with the earth is short circuited). The electrical fields due to the wave induce charges in the earth's surface as shown in Fig. 16.11. As the wave travels, the induced charges in the earth or ground also travel along with it. This constitutes a current in the earth's surface. As the *ground wave* passes over the surface of the earth, it is weakened as a result of energy absorbed by the earth. Due to these

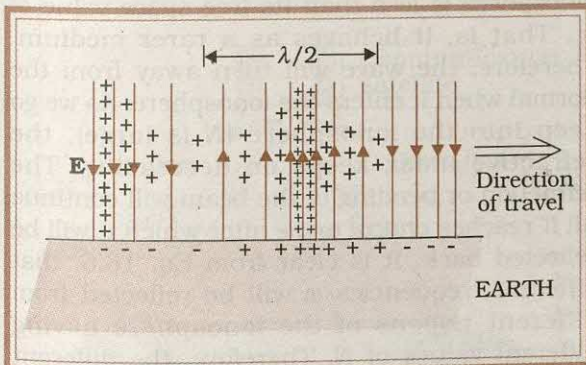


Fig. 16.11 Vertically polarised wave travelling over the surface of the earth. The solid lines represent the electric field (**E**) of the electromagnetic wave.

losses the *ground waves* are not suited for very long range communication. Further, these losses are frequency dependent and are higher for high frequencies. Hence, ground wave propagation can be sustained only at low frequencies (~ 500 kHz to 1500 kHz) or for radio broadcast at long wavelengths. The higher frequencies (≥ 1 MHz) propagate through the following two modes, viz., space wave and sky wave propagations.

16.7.2 Space Wave Propagation or Tropospheric Wave Propagation.

The transmitted waves, travelling in a *straight line*, directly reach the receiver end and are then picked up by the receiving antenna as shown in Fig. 16.12. It can be seen that due to the finite curvature of the earth, such waves cannot be seen beyond the *tangent* points R_1 and R_2 . The effective reception range of the broadcast is essentially the region from R_1 to R_2 which is covered by the *line of sight* in a conventional sense. Hence, sometimes this mode of communication is termed as **line of sight communication**.

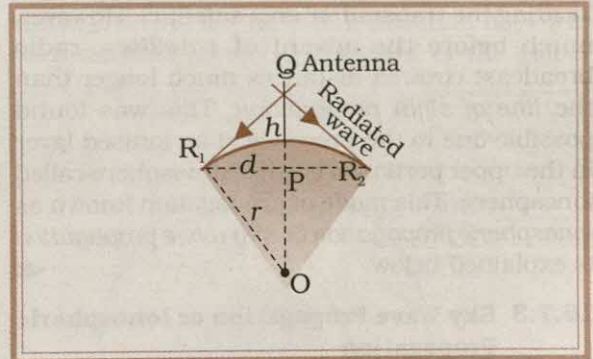


Fig. 16.12 Ray path of transmitted waves following space-wave (or line of sight) mode of propagation. The transmitter is located at the ground on a tall tower.

The range R_1R_2 (or $PR_1 = d$ which is half the total range) can be easily calculated by geometrical consideration. Suppose height of the tower is h and the radius of earth is r (that is $OR_1 = OR_2 = OP = r$). In the right-angled triangles OQR_1 , QPR_1 we have

$$OQ^2 = QR_1^2 + OR_1^2$$

$$QR_1^2 = h^2 + d^2$$

Then,

$$(r + h)^2 = r^2 + d^2 + h^2$$

$$r^2 + h^2 + 2rh = r^2 + d^2 + h^2$$

or $d^2 = 2rh$

$$d = \sqrt{2rh} \quad (16.4)$$

This distance is of the order of 40 km.

Example 16.2 A TV tower has a height of 75 m. What is the maximum distance and area upto which this TV transmission can be received. Take radius of the earth as 6.4×10^6 m.

Answer

$$d = \sqrt{2rh} \quad (16.4)$$

$$= \sqrt{2 \times 6.4 \times 10^6 \times 75}$$

$$= 3.1 \times 10^4$$

$$= 31 \text{ km.}$$

$$\begin{aligned} \text{Area covered} &= \pi d^2 = 2\pi rh \\ &= 3018 \text{ km}^2 \end{aligned}$$

If one wishes to send signals at far away stations, then either *Repeater* transmitting stations are necessary or h is increased (by locating the transmitter on a satellite). However, much before the advent of *satellites*, radio broadcast covered distances much longer than the *line of sight propagation*. This was found possible due to the presence of an ionised layer in the upper portion of earth's atmosphere called **ionosphere**. This mode of propagation known as *ionospheric propagation* or *sky wave propagation* is explained below.

16.7.3 Sky Wave Propagation or Ionospheric Propagation

This mode of propagation is shown in Fig. 16.13. A transmitted wave going up in the sky is reflected back from the ionised region of the earth's atmosphere, the **ionosphere**. The UV and other high energy radiations coming from Sun are absorbed by air molecules which get ionised and form an ionised layer of electrons and ions around the earth. The ionosphere extends from a height of ~ 80 to 300 km above the earth's surface.

Without going into the mathematical details, we can easily understand the reflection of electromagnetic waves from the ionosphere. The oscillating electric field of electromagnetic wave (frequency ω) changes the velocity of the

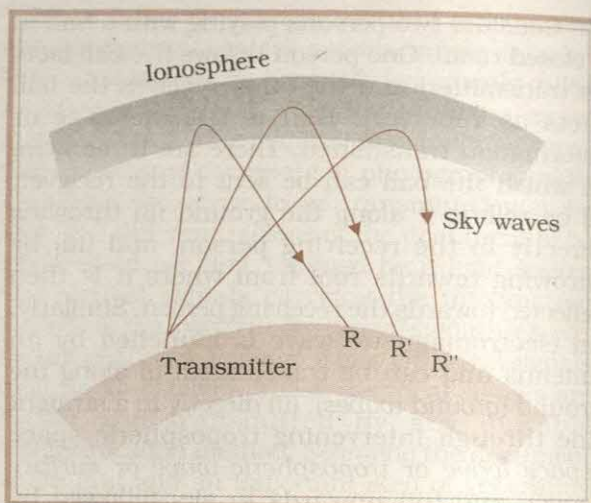


Fig. 16.13 Ionospheric or Sky wave propagation.

electrons in the ionosphere (negligible change for ions because they are heavy and electromagnetic wave field is weak). This changes the effective dielectric constant (ϵ'), and hence the refractive index (n'), as compared to free space values of ϵ_0 and n_0 . The effective dielectric constant ϵ' and the corresponding refractive index n' are related to ϵ_0 and n_0 as

$$\begin{aligned} n' &= \sqrt{\epsilon' \mu_0} \\ &= \sqrt{\epsilon_0 \mu_0 [1 - (Ne^2 / \epsilon_0 m \omega^2)]} \end{aligned} \quad (16.5)$$

$$\text{or } n' = n_0 [1 - (Ne^2 / \epsilon_0 m \omega^2)]^{1/2} \quad (16.6)$$

where e is the electronic charge, m is the mass of the electron and N is the electron density in the ionosphere.

It is clear that the refractive index of ionosphere is less than its free space value of n_0 . That is, it behaves as a rarer medium. Therefore, the wave will turn away from the normal when it enters the ionosphere. As we go deep into the ionosphere (N is large), the refractive index keeps on decreasing. The refraction or bending of the beam will continue till it reaches *critical angle* after which it will be reflected back. It is clear from Eq. 16.6, that different frequencies ω will be reflected from different regions of the ionosphere having different values of N . Therefore, the different points on earth receive signals reflected from different depths of the ionosphere. If the frequency ω is too high, then after a certain value, the electron density N may never be so

high as to produce enough bending for attainment of critical angle or condition of reflection. This is called *critical frequency*. If maximum electron density of the ionosphere is N_{\max} per m^3 , then the critical frequency f_c is approximately given by:

$$f_c \approx 9 (N_{\max})^{1/2} \quad (16.7)$$

The f_c ranges approximately from 5 to 10 MHz. Frequencies higher than this cross the ionosphere and do not return back to the earth.

Long distance communication beyond 10 to 20 MHz was not possible before 1960 because all the three modes of communication failed (*ground waves* due to conduction losses, *space wave* due to limited line of sight, *sky wave* due to the penetration of the ionosphere by the high frequencies beyond f_c). Now, this is possible with the advent of the new concept of *satellite communication*.

16.8 SATELLITE COMMUNICATION

Present day requirements of information technology demand a very large number of communication channels and hence large frequency bandwidth. This is only possible if the communication frequency is high. We have already seen above that the waves of frequencies $f > 30$ MHz cannot propagate by conventional modes. A new concept of communication, the communication through a satellite, has revolutionised communication technology.

The basic principle of satellite communication is schematically shown in Fig. 16.14. A communication satellite is a spacecraft placed in

an orbit around the earth which carries a transmitting and a receiving equipment (termed as *Radio Transponder*). The transmitted signal is **UP-LINKED** and received by the satellite station which **DOWN-LINKS** it with the ground station through its transmitter. To avoid confusion, the **up-link** and **down-link** frequencies are kept different. Both these frequencies being in the UHF/microwave regions, can cross the ionosphere and reach the satellite located well above the ionosphere; in fact, the height of geostationary satellite is $\sim 36,000$ km.

The most important consideration in designing a satellite communication system is to choose the orbit in which the satellite is to be placed. For steady reliable transmission and reception, it is preferred that the satellite should be geostationary. A geostationary satellite is one that appears to be stationary relative to the earth. It has a circular orbit lying in the equatorial plane of the earth (inclination 0°) at an approximate height of 36,000 km. If we use three geostationary satellites placed at the vertices of an equilateral triangle as shown in Fig. 16.15(a), the entire earth (globe) can be linked/covered by the communication network, each satellite covering one-third of the globe.

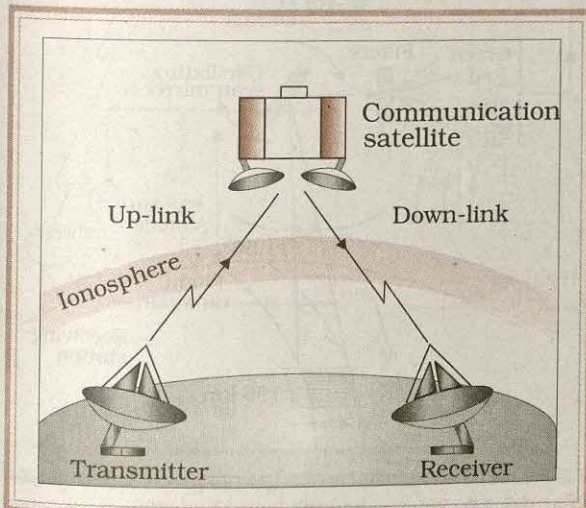


Fig. 16.14 Principle of satellite communication.

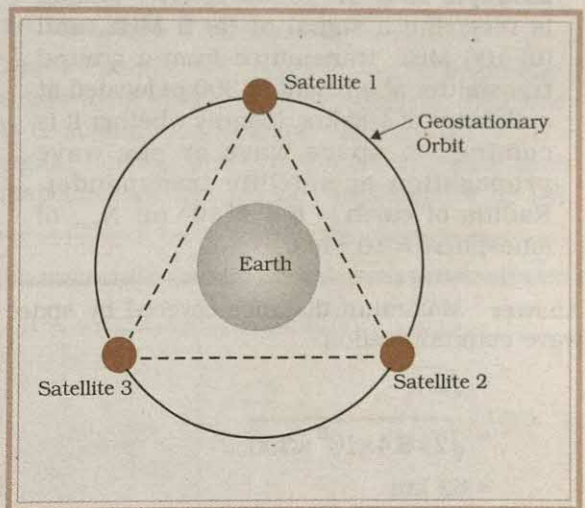


Fig. 16.15(a) Three communication satellites placed on equilateral points on geostationary orbit for global communication coverage.

Most satellites are located in the geostationary equatorial orbit. However, there are two more orbits which are being used for communication [Fig. 16.15(b)]. These are

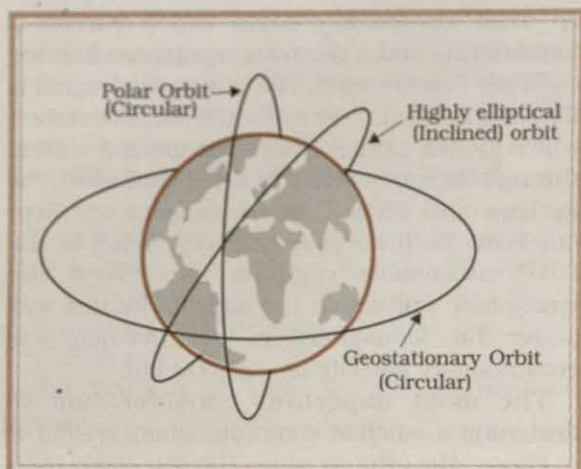


Fig. 16.15(b) A schematic diagram of various satellite orbits used in satellite communication.

(a) *polar circular orbit* which is nearer to earth (1000 km away) as compared to the geostationary orbit. It passes over, or very close to the poles (i.e., inclination $\approx 90^\circ$). (b) *Inclined highly elliptical orbit* which is used for communications in regions of high latitudes. The preferred inclination is $\approx 63^\circ$ and hence such orbits are referred to as being in **63° slot**.

Example 16.3 A ground receiver station is receiving a signal at (a) 5 MHz, and (b) 100 MHz, transmitted from a ground transmitter at a height of 300 m located at a distance of 100 km. Identify whether it is coming via space wave or sky wave propagation or satellite transponder. Radius of earth $\approx 6.4 \times 10^6$ m; N_{\max} of ionosphere $= 10^{12} \text{ m}^{-3}$.

Answer Maximum distance covered by space wave communication

$$\begin{aligned}
 &= \sqrt{2rh} \\
 &= \sqrt{2 \times 6.4 \times 10^6 \times 300} \\
 &= 62 \text{ km}
 \end{aligned}$$

Since receiver-transmitter distance is 100 km, this is ruled out both for 'a' and 'b'.

Further, f_c for ionospheric propagation is

$$\begin{aligned}
 f_c &= 9 (N_{\max})^{1/2} \\
 &= 9 \times (10^{12})^{1/2} \\
 &= 9 \text{ MHz}
 \end{aligned}$$

So, the 'a' signal of 5 MHz ($< f_c$) comes via

ionospheric mode while the 'b' signal of 100 MHz comes via the satellite mode.

16.9 REMOTE SENSING: AN APPLICATION OF SATELLITE COMMUNICATION

Remote sensing is the science and art of obtaining information about an object, area, or phenomenon, acquired by a sensor that is not in direct contact with the target of investigation. Any photography is a kind of remote sensing. If we want to cover large areas for which information is required, we have to take photograph from larger distances. Aerial photography was first introduced in World War I for military uses and later extended to areas such as archaeology, resource surveys, town and country planning. Now the more advanced technology of **satellite imagery** has emerged.

A satellite equipped with appropriate sensors to acquire data can be placed in an orbit around the earth at any height having a period of revolution. It takes photographs or collects any other information desired and transmits it back to an earth station. This is known as remote sensing and is schematically illustrated in Fig. 16.16. This figure illustrates the schematics of **LANDSAT** series of satellites designed for land-use applications. These satellites move in near-polar orbits at an altitude of 918 km. The instrument on board which, over the years, has provided most of the data is a **multispectral scanner**. This instrument scans a swath of width

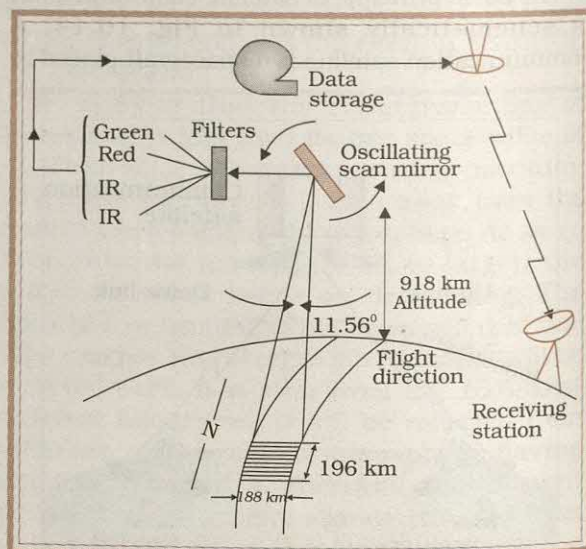


Fig. 16.16 Schematic illustration of multispectral scanner remote sensing system.

about 185 km as shown in the Fig. 16.16. The satellite travels in a direction slightly west of south and make **passes** over the desired region many times a day. Thus, different parts of earth are scanned. The data obtained is transmitted to the receiving station on the earth.

Taking photograph of any object relies on the reflected wave from the object as we use visible light in normal photography. In principle, waves of any wavelength in the electromagnetic spectrum can be used for this purpose by employing suitable sensors. However, selection of wavelength of the radiation depends on the effect of atmosphere including ionosphere and the nature of objects to be scanned.

The visible and near infrared spectral bands (which can be displayed as colours) are chosen to amplify (or to separate) specific earth features such as vegetation and water. This is because the reflectance from different objects is different. In many cases, the data from the thermal infrared bands are of high interest, particularly due to the fact that the thermal infrared data is a measure of surface temperature and can also be obtained at night. Microwave data are of particular relevance for certain hydrological variables such as soil moisture and precipitation. They can also be obtained at night and are not restricted to cloud-free conditions.

Some applications of remote sensing include meteorology (development of weather systems and weather forecasting), climatology (monitoring climate changes), oceanography (sea surface temperatures, mapping of sea-ice and

oil pollution monitoring), and coastal studies (sewage, industrial waste and pollution monitoring). Archaeology, geological surveys, water resource surveys, urban land use surveys, agriculture and forestry, and natural disaster constitute other applications of remote sensing.

16.10 LINE COMMUNICATION

In the space communication, we have seen that there was no physical point-to-point contact between the receiver and the transmitter. However, the simplest and the oldest point-to-point communication mode is communication through **transmission lines** or **wires** as in earlier telephone and telegraph links. These are still used though significant changes in the design of such lines have occurred. The principal types of such communication channels are:

- (i) **Two wire transmission lines**
- (ii) **Coaxial cables**
- (iii) **Optical fibres**

Communication channels (i) and (ii) are employed for AF to UHF regions, and (iii) is used for optical frequencies.

The type (i) and (ii) transmission lines are briefly described in the following sub-sections while the type (iii) transmission line is described in Section 16.12.

16.10.1 Two Wire Transmission Line

You are familiar with such lines and might have seen wire of telephone lines along the road side. It is schematically shown in Fig. 16.17. The signal flowing through the wires create their respective electric (E) and magnetic fields (H).

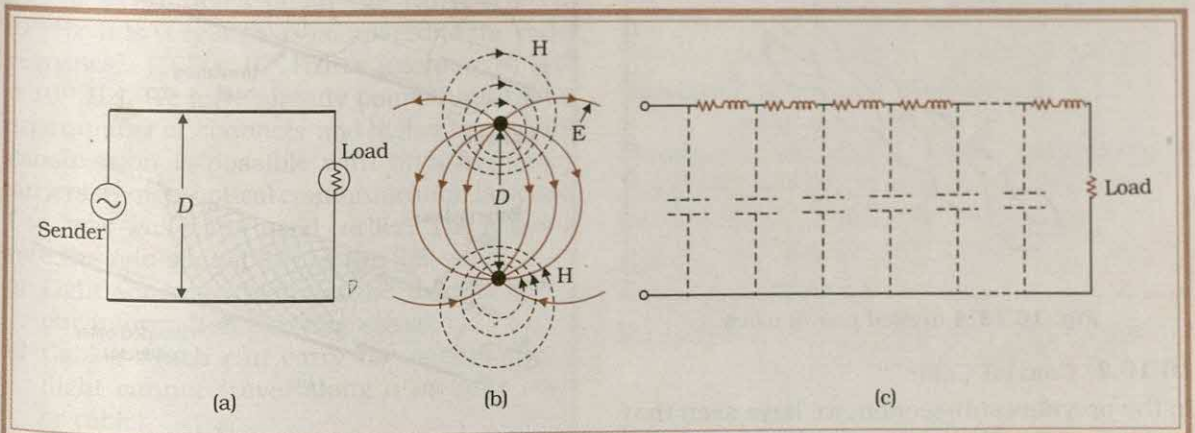


Fig. 16.17 Description of a two wire wave transmission line: (a) a typical line; (b) electric (solid lines) and magnetic (dashed lines) fields due to the signal in the transmission line; and (c) equivalent circuit of transmission line geometry shown in (a). The impedances are distributed throughout the line.

It has been shown that the high frequency signals travel as wave in the transmission line. Each portion of the transmission line can be considered as a *small* inductor, resistor and capacitor as shown in Fig. 16.17(c). Such inductors, resistors and capacitors are *distributed* throughout the transmission line. As a result, each length of the transmission line has a **characteristic impedance**. When the impedance of the detecting device at the receiver (load) is matched (i.e. equal) to the characteristic impedance, then maximum power is transferred to the load otherwise load matching is necessary. You might have seen telephone line-man **matching** the load when the reception in your telephone is not good.

As the frequencies being transmitted increase towards RF or more, the impedance effects ωL

or $\frac{1}{\omega C}$ or even R (R increases with frequencies

due to *Skin effect*) become significant. So, such transmission lines become very lossy. Even more serious problem arises because at high radio frequencies the transmission lines start **radiating**. Additionally, electromagnetic interference from other radiating sources also occurs. The latter can be somewhat reduced if we use a twisted wire geometry shown in Fig. 16.18. Normally, for commercial applications, a number of such twisted pairs are bundled together into a single cable. Such cables are being extensively used in our local telephone or computer networking.

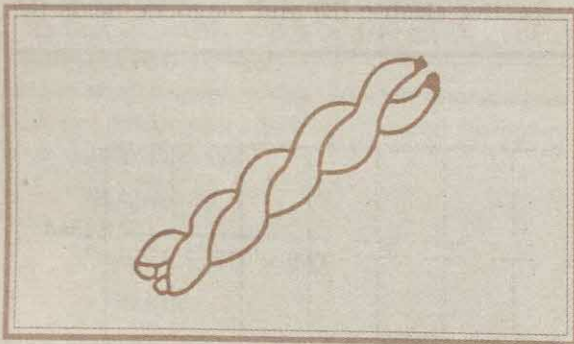


Fig. 16.18 A twisted pair of wires.

16.10.2 Coaxial Cable

In the previous sub-section, we have seen that a two wire transmission line at high frequencies suffer from the disadvantage of being prone to electromagnetic interference

and also become sources of radiation themselves. This can be avoided by use of shielded coaxial cables. A coaxial cable consists of a hollow outer cylindrical conductor which surrounds a single inner conductor kept separated from each by an insulator. Schematically, it is shown in Fig. 16.19(a). The electromagnetic fields at its cross-section are shown in Fig. 16.19(b). The outer conductor acts as the shield and minimises interference.

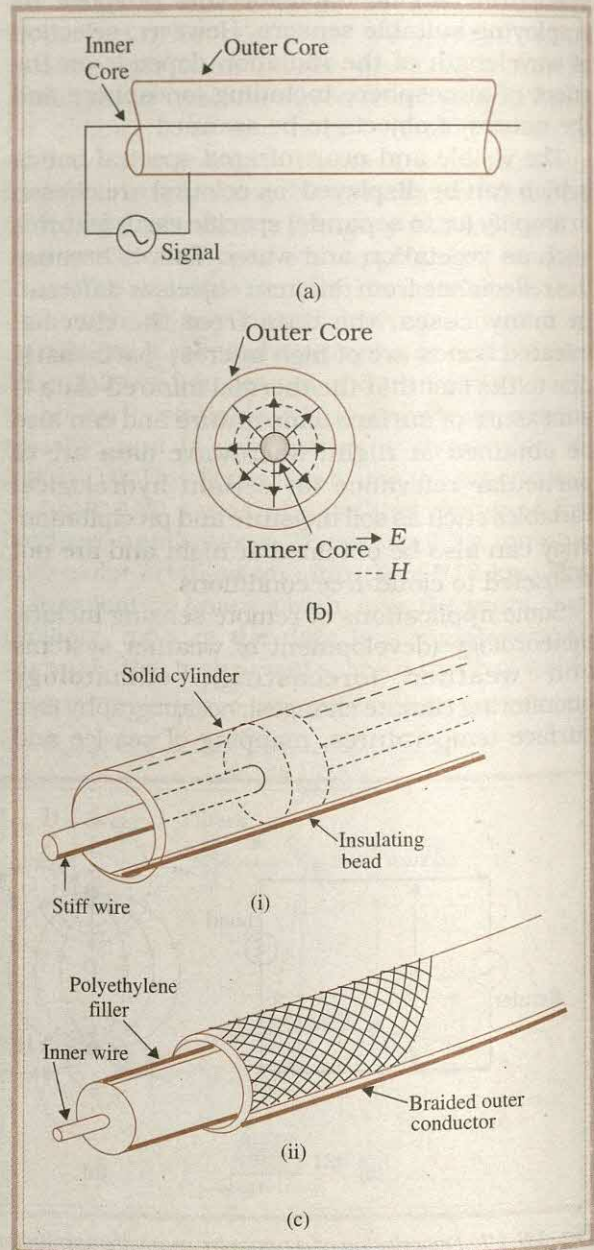


Fig. 16.19 Description of a coaxial cable.

Some constructional details of a coaxial cable are shown in Fig. 16.19(c). The inner conductor is made up of a copper wire and the outer conductor can be either a solid or a braided mesh of fine wires. The inner conductor is held at the centre by a solid dielectric (insulating) material acting as a *spacer*. Different kinds of dielectric materials, such as Teflon and Polyethylene, are used depending upon the frequency and power to be transmitted through the cable. The outer conductor is normally connected to ground and thus it provides an electrical shield to the signals carried by the central conductor. The outer conductor is externally covered with a polymer jacket for protection. The closed structure prevents inner copper wire or core from radiating signal power, thus reducing losses.

In the transmission of power through coaxial cable, the dielectric medium used as spacer plays an important role. These dielectrics are good insulators only for dc or low frequencies. As the frequency increases, there is significant *dielectric loss* and can sometimes become quite high after a certain frequency. So, it puts an upper limit (approximately 20 MHz) upto which a coaxial cable can be effectively used. Very often, repeater stations become necessary if the distances are more than a few kilometers.

16.11 OPTICAL COMMUNICATION

The name implies that the communication is through carrier optical signals. Easily accessible optical frequencies lie in the range 10^{12} to 10^{16} Hz. It is very high as compared to the radio frequencies (10^6 to 10^8 Hz) or microwaves (10^9 to 10^{11} Hz). We have already pointed out that a large number of *channels* and higher bandwidth transmission is possible with high frequency carriers. Hence, optical communication is better. If so, why was it not used earlier? The reasons were the non-availability of the following:

- Light sources which can be modulated by our information carrying signal.
- Cables which can carry the optical signal (light cannot travel along a metallic wire or cable).

With the recent availability of *Optical fibres* of small diameters ($\sim 10^{-5}$ m) (through which light can be transmitted), semiconducting light

sources (LED, diode laser) and detectors (photodiodes, Avalanche photodiode), the optical communication technology is developing very rapidly.

There are some inherent advantages of optical communication over the conventional two-wire or cable electronic communication systems. Some of these are:

- Wide channel bandwidth and large channel carrying capacity** because of the use of higher frequencies $\sim 10^{14}$ Hz as compared to the electronic communication links.
- Low transmission losses**. In optical fibres losses per km are less.
- Signal security and not accessible to interference**. You will see that the optical signal is confined to the inside of fibre and cannot be tempered easily. So secret information like banking, defence etc. is more secure.

Example 16.4 Consider an optical communication system operating at $\lambda \sim 800$ nm. Suppose, only 1% of the optical source frequency is the available channel bandwidth for optical communication. How many channels can be accommodated for transmitting: (a) audio-signals requiring a bandwidth of 8 kHz, and (b) Video TV signals requiring an approximate bandwidth of 4.5 MHz.


Answer Optical source frequency, $f = c/\lambda$
 $= 3 \times 10^8 / (800 \times 10^{-9})$
 $= 3.8 \times 10^{14}$ Hz

Bandwidth of channel (1% of above)
 $= 3.8 \times 10^{12}$ Hz

Number of channels = (Total bandwidth of channel) / (Bandwidth needed per channel)

(a) Number of channels for audio signal
 $= (3.8 \times 10^{12}) / (8 \times 10^3)$
 $\sim 4.8 \times 10^8$

(b) Number of channels for video TV signal
 $= (3.8 \times 10^{12}) / (4.5 \times 10^6)$
 $\sim 8.4 \times 10^5$

You can see that the number of channels is very large which can be simultaneously transmitted. 

16.11.1 Basic Optical Communication Link

The simplest optical transmission link is a point-to-point link having a transmitter at one end and a receiver at the other end as shown in Fig. 16.20.

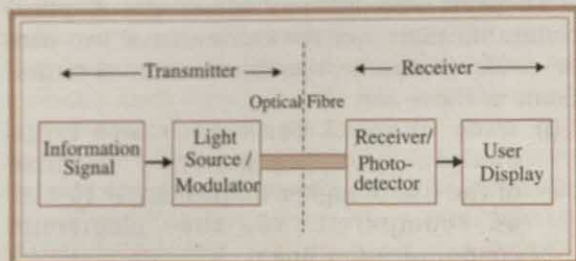


Fig. 16.20 Point-to-point optical communication link.

The components to be employed in any optical communication system depend upon the system requirements like desired transmission distance, channel bandwidth, cost etc. For example, for a short distance transmission you can use LED or diode laser of wavelength (λ) about 800 nm to 900 nm. For large distances, we have to use a wavelength which is attenuated less, say, $\lambda \sim 1300$ to 1550 nm. Similarly, for obtaining strong optical signals at the receiver end, a p-n photodiode may be sufficient but for detecting weak optical signal you may prefer to use highly sensitive (but expensive) avalanche photodiode.

The three most important components of an optical communication link are:

- (i) **Optical source and modulator**
- (ii) **Optical signal detector or photodetector**
- (iii) **Optical fibre cable through which optical signal is transmitted**

Basics of the first two of the above have already been discussed in Chapter 15 (Section 15.8) and hence, these would only be briefly dealt with in this Chapter. The last component, namely optical fibre, is the key to the rapid growth and success of the modern optical communication system. This would be treated in some detail in Section 16.11.2 without going into the mathematical formulations.

16.11.2 Optical Sources for Communication Links

Any intense monochromatic source of light, in principle, is suitable. However, semiconductor Light Emitting Diodes (LED) and diode lasers are preferred light sources for the following reasons:

- (i) These give adequate power.
- (ii) Diode laser light is monochromatic and coherent. Therefore, they can easily be obtained as parallel beams spread over a narrow cone or angle. These are preferred for very large distance transmission.
- (iii) LED also gives near monochromatic light (not as much as in diode laser). This near monochromaticity is sufficient for applications over small distances. LED is a low cost device as compared to diode laser.
- (iv) The dimensional characteristics of both diode laser and LED are compatible with those of the optical fibre, which means that coupling of fibres with light source will be easier.
- (v) The light from LED or diode laser can be varied or modulated by varying the applied voltage or current through the device which is essentially the information to be transmitted.

The construction and working of LED and diode laser has already been discussed in Chapter 15 (Section 15.8). From operational point of view for optical communication, the following significant points have to be considered:

- (a) **Light modulation by the relevant information signal:** The information (like speech, music, digital code etc.) can generally be made available in the form of electrical signal. We know that the light intensity in LED or diode laser varies with the applied voltage or current through the device. Hence, the applied information signal voltage produces a modulated light signal. For faithful modulation, the LED or diode laser is biased near its threshold voltage so that small variations in the information signal voltage produce the desired light intensity modulation.
- (b) **Thermal stability:** The frequency and intensity of light is sensitive to temperature changes (which has to be avoided) since the band gap and electron density in different energy states of a semiconductor varies with temperature. Therefore, if LED or diode laser has to be used in an environment where temperature changes are large, a suitable arrangement for temperature control and/or cooling is desirable particularly for diode laser.

16.11.3 Optical Detectors

The optical signal reaching the receiving end has to be detected by a **detector** which converts light into electrical signal so that the transmitted information may be decoded. The optical detector should have:

- (i) size compatible with the fibre,
- (ii) high sensitivity at the desired optical wavelength, and
- (iii) high response time for fast speed data transmission/reception.

Semiconductor based photodetectors are used because they fulfill all the criteria required for the detector of an optical communication system.

The simple **photodiode** which is generally used as an optical detector is described in Chapter 15 (Section 15.8). When light photons with energy, $h\nu > E_g$ (band gap of semiconductor) fall on a photodiode, electron-hole pairs are generated which are separated by the reverse bias applied to it. This gives us a current. In general, the efficiency of generating electron-hole pairs in a photodiode decides how good that detector is.

Example 16.5 A photodetector is made from a semiconductor $\text{In}_{0.53}\text{Ga}_{0.47}\text{As}$ with $E_g = 0.73 \text{ eV}$. What is the maximum wavelength, which it can detect?

Answer Limiting value of $h\nu$ is E_g , such that

$$h\nu = \frac{hc}{\lambda} = E_g$$

$$\text{or } \lambda = \frac{hc}{E_g}$$

$$= \frac{6.63 \times 10^{-34} \text{ J s} \times 3 \times 10^8 \text{ m s}^{-1}}{0.73 \times 1.6 \times 10^{-19} \text{ J}} \\ = 1703 \text{ nm}$$

16.12 OPTICAL FIBRE

In a point-to-point communication link using electrical signal, we use a conducting wire or a cable and the signal current is *guided* through the conductor cable wire. If we wish to use optical signals, then we need to have some cable or wire-type thing along which the light could travel in the desired (or guided) direction. Light cannot pass through metal but can pass through

a transparent glass, polymer or dielectric. Therefore, optical transmission line (called optical fibre) is made from any of these materials. How can light be restricted within the glass fibre? The optical fibres make use of the principle of total internal reflection (Chapter 10).

The structure of an optical fibre, mechanism of propagation of light along it, methods and materials used for fabricating an optical fibre are discussed below:

A typical optical fibre, as shown in Fig. 16.21(a), consists of:

- (i) **Core** of glass/silica/plastic with approximate diameter of 10 to 100 μm with refractive index n_1 . The plastic core has high loss and hence glass cores are preferred.
- (ii) The core is surrounded by a glass or a plastic **cladding** with refractive index n_2 such that $n_2 < n_1$. The difference $\Delta n = n_1 - n_2$ is typically very small of the order of 10^{-3} . The refractive index of cladding can change either abruptly from n_1 to n_2 (a *step-index fibre*, [Fig 16.21(b)] or gradually (*graded-index fibre*) [Fig. 16.21(c)].
- (iii) For providing safety and strength, a **buffer** plastic coating or housing encapsulates the **core-cladding** of the fibres.

Now we try to understand how light can travel and remain confined to the fibre. Let us take a general case of light incident on the **core** of optical fibre at any angle ϕ_i as shown in Fig. 16.22(a). A part will be reflected and a part will be refracted at an angle ϕ_r , running away from the normal (since $n_2 < n_1$). The angle ϕ_i (and hence θ_i and θ_r) are related by Snell's law:

$$n_1 \sin \phi_i = n_2 \sin \phi_r$$

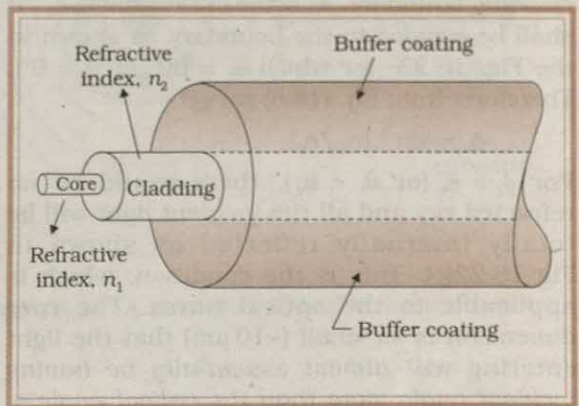


Fig. 16.21(a) A schematic diagram of single optical fibre structure.

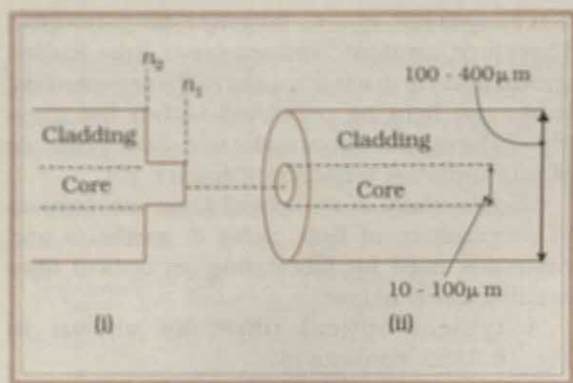


Fig. 16.21(b) Abruptly changing refractive indices at the core-cladding boundary 'a step-index fibre'.

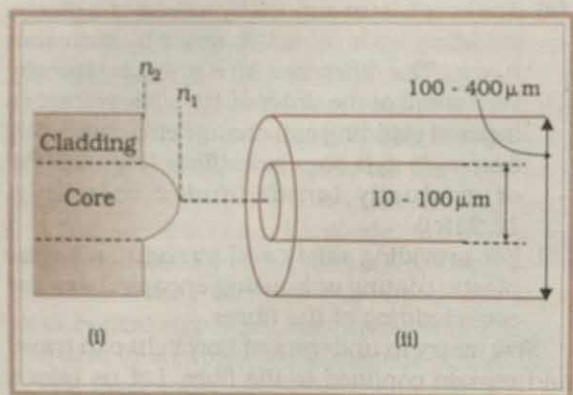


Fig. 16.21(c) Gradually changing refractive indices from core to cladding : graded-index fiber.

$$\text{or } n_1 \cos \theta_i = n_2 \cos \theta_r \quad (16.8)$$

Suppose ϕ_i is increased (or θ_i is decreased), the refracted ray will come nearer to the core-cladding boundary. At some critical angle ϕ_c , it shall be parallel to the boundary as shown in the Fig. 16.23 for which $\phi_r = 90^\circ$ or $\theta_r = 0^\circ$. Therefore, from Eq. (16.8) we get

$$\phi_c = \sin^{-1} (n_2 / n_1).$$

For $\phi_i > \phi_c$ (or $\theta_i < \theta_c$), there would be no refracted ray and all the incident light will be totally internally reflected as shown in Fig 16.22(c). This is the condition, which is applicable to the optical fibres. The core dimension is so small ($\sim 10 \mu\text{m}$) that the light entering will almost essentially be having incident angle more than the critical angle ϕ_c and will suffer *total internal reflection* at the core-cladding boundary. Such successive total

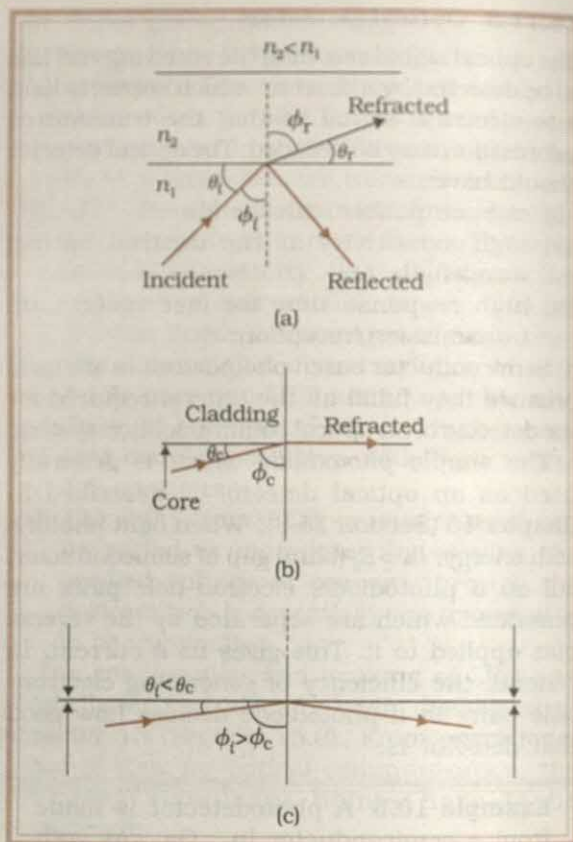


Fig. 16.22 Light ray paths near core-cladding boundary of optical fibre (a) ϕ_i is small (b) $\phi_i = \phi_c$: critical angle (c) $\phi_i > \phi_c$ or $\theta_i < \theta_c$.

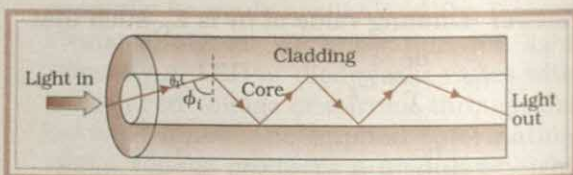


Fig. 16.23 Ray confinement in actual optical fibre.

reflections at opposite boundaries will *confine* the light to the core as shown in Fig. 16.23. Thus, the light travels in the core in a *guided* manner and hence optical fibre is sometimes called as **optical waveguide**.

You must have realised from the above discussion that for ensuring incident angle $\phi_i > \phi_c$, the core (and cladding) diameters have to be small. Making such fibres is technologically *tricky* as the optical fibres have the speciality that the inner core glass is different (refractive index n_1) than the cladding glass (refractive index n_2). The principle of fabricating optical fibres is given in the box below.

PRINCIPLE OF FABRICATING OPTICAL FIBRES

You might have seen that if we *soften* (nearly melt) glass and pull it, then we get thin fibres. For obtaining optical fibres two steps are involved.

Step 1: Preformed glass rods (1-2 cm diameter) are first prepared with different inner and outer glass compositions. One simple method for obtaining **preform** is shown in Fig. 16.24. A core glass rod is taken which is being constantly rotated. SiCl_4 or GeCl_4 vapours burn in a burner and give oxidised SiO_2 or GeO_2 powder which gets vitrified and modifies the glass property at the outer core of the original glass rod. Thus, we have a **preform** glass rod in which inner core glass is different from the outer cover or clad glass.

Step 2: The **preformed** glass rod is softened in a furnace and fibres are drawn as shown in Fig. 16.25. Finally, as a next step the plastic buffer coating is done to obtain finished optical fibre. For use in telecommunication system many fibres are usually incorporated into a cable structure as shown in Fig. 16.26.

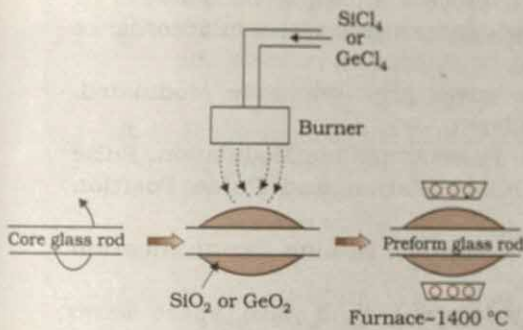


Fig. 16.24 Steps in obtaining glass-rod preform for use in drawing optical fiber.

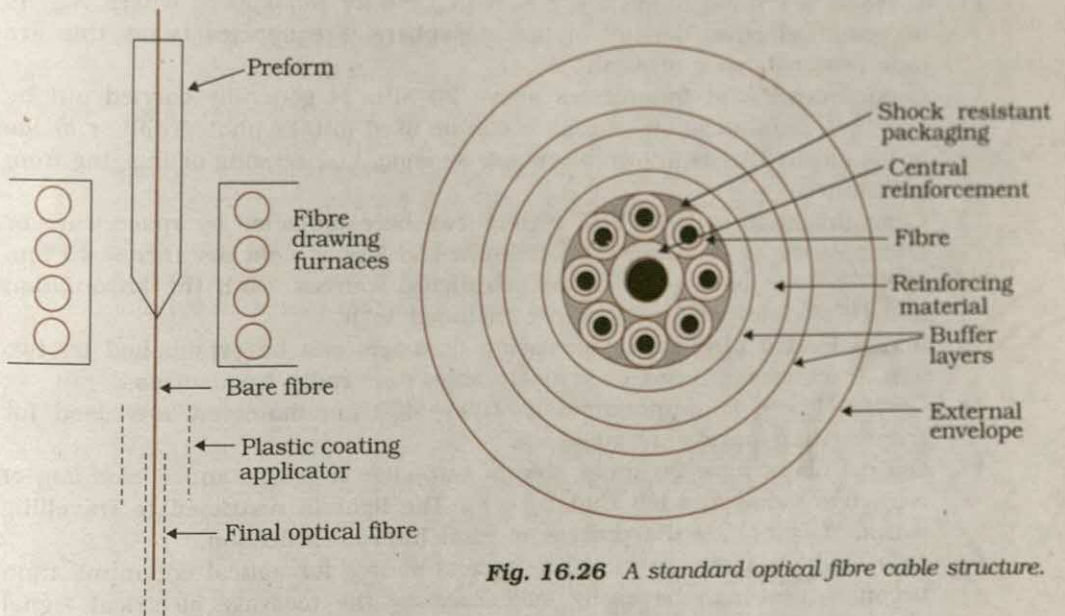


Fig. 16.26 A standard optical fibre cable structure.

Fig. 16.25 Schematic representation of drawing optical fibre from glass-preform.

SUMMARY

1. Electronic communication refers to the faithful transfer of information or message (available in the form of electrical voltage or current) from one point to another point.
2. A message signal is defined as a single valued function of time (that conveys the information) and which, at every moment of time has a unique value.
3. Transmitter, transmission channel and receiver are three basic units of a communication system.
4. Two important forms of communication systems are: Analog and Digital. The information to be transmitted is generally in continuous wave form for the former while for the latter it has only *discrete* or *quantised* signal levels.
5. Low frequencies cannot be transmitted to long distances. These are loaded on a high frequency carrier signal by a process known as *modulation*.
6. In modulation, some characteristic of the carrier signal varies in accordance with the modulating or message signal.
7. Types of modulated sinusoidal carrier waves are: Amplitude Modulated, Frequency Modulated and Phase Modulated.
8. Pulse modulation could be classified as: Pulse Amplitude Modulation, Pulse Duration Modulation or Pulse Width Modulation and Pulse Position Modulation.
9. There are two types of transmission channels: (i) Line communication and (ii) Free space communication.
10. Different space wave propagation methods are: ground wave, space wave, or tropospheric wave having line of sight propagation, ionospheric or sky wave communication, satellite communication.
11. There is a critical frequency $f_c = 9 (N_{\max})^{1/2}$ for ionosphere where N_{\max} is maximum electron density in the ionosphere. Frequencies below this are only reflected back to earth.
12. Communication at frequencies above 20 MHz is generally carried out by satellite communication. Satellites can be used to take photograph or *image* of the earth. This is known as *remote sensing*, i.e., sensing or imaging from a distance.
13. Communication in UHF/VHF regions can be established by space wave or tropospheric wave. However, it is limited to *line of sight* distance ≈ 40 km.
14. MODEMS are used to interface two digital sources. Both the MODulation and DEModulation functions are included in it.
15. Actual guided physical transmission channels can be established by two wire transmission lines or coaxial cables for radio frequencies.
16. Optical fibres (of approximately 100 – 400 μm diameter) are used for transmitting optical frequencies.
17. Optical fibers have an inner *core* of refractive index n_1 and a *cladding* of refractive index n_2 such that $n_2 < n_1$. The light is restricted to travelling within the fibre by the process of total internal reflection.
18. LED or diode lasers are used as optical source for optical communication because these can be easily *modulated* by the message electrical signal voltage.
19. Semiconducting photodiodes are used as detectors in optical communication network.

EXERCISES

- 16.1** At which of the following frequency/frequencies the communication will not be reliable for a receiver situated beyond horizon:
- 100 kHz
 - 1 MHz
 - 1GHz
 - 100 GHz
- 16.2** Modulation is used to
- reduce the bandwidth used.
 - separate the transmissions of different users.
 - ensure that intelligence may be transmitted to long distances.
 - allow the use of practical antennas.
- 16.3** AM is used for broadcasting because
- it is more noise immune than other modulation systems.
 - it requires less transmitting power compared with other systems.
 - its use avoids receiver complexity.
 - no other modulation system can provide the necessary bandwidth faithful transmission.
- 16.4** Frequencies in the UHF range normally propagate by means of
- ground waves.
 - sky waves.
 - surface waves.
 - space waves.
- 16.5** Digital signals (i) do not provide a continuous set of values, (ii) represent values as discrete steps, (iii) can utilize only binary system, and (iv) can utilize decimal as well as binary system. Which of the following options is **True**:
- Only (i) and (ii).
 - Only (ii) and (iii).
 - Only (i), (ii) and (iii), but not (iv).
 - All the above (i) to (iv).
- 16.6** A microwave telephone link operating at the central frequency of 10 GHz has been established. If 2% of this is available for microwave communication channel, then how many telephones channel can be simultaneously granted if each telephone is allotted a bandwidth of 8 kHz.
- 16.7** The core of an optical fibre is made of glass with refractive index equal to 1.55 and clad has refractive index 1.51. Calculate (a) the critical angle for total internal reflection, and (b) maximum acceptance angle at the air-core interface.
- 16.8** You are given three semiconductors: A, B and C with respective bandgaps of 3 eV, 2 eV and 1 eV for use in a photodetector to detect $\lambda = 1400$ nm. Select the suitable semiconductor. Give reasons.
- 16.9** Frequencies higher than 10 MHz are found not to be reflected by the ionosphere on a particular day at a place. Calculate the maximum electron density of the ionosphere.
- 16.10** A radar has a power of 1 kW and is operating at a frequency of 10 GHz. It is located on a steep mountain top of height 600 m. What is the maximum distance upto which it can detect an object located on the surrounding earth's surface?
- 16.11** In the Exercise 16.10, suppose the antenna at the mountain top is mounted on a tower of height 100 m and the Radar power is doubled to 2 kW. What is the maximum distance which it can cover now?

- 16.12** On a particular day, the maximum frequency reflected from the ionosphere is 10 MHz. On another day, it was found to increase to 11 MHz. Calculate the ratio of the maximum electron densities of the ionosphere on the two days. Point out a plausible explanation for this.

ADDITIONAL EXERCISES

- 16.13** The intensity of a light pulse travelling along a fibre decreases exponentially with distance according to the equation

$$I = I_0 e^{-\alpha x}$$

where I_0 is the intensity at $x = 0$ and α is the attenuation constant. Show that

the intensity reduces by 50% after a distance of $\frac{\ln 2}{\alpha}$.

- 16.14** Light from a source located in a medium (refractive index = n_0) enters an optical fibre with core refractive index of n_1 and clad refractive index of n_2 as shown in the Fig. 16.27.

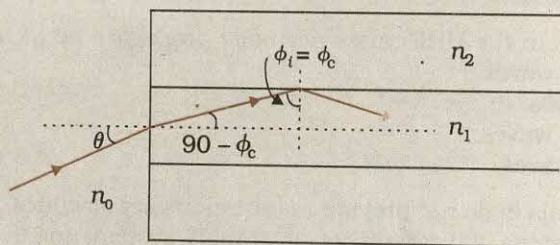


Fig. 16.27

Show that the maximum value of internal incidence angle θ_0 (called maximum acceptance angle) which would undergo total internal reflection in the fibre (and will be able to travel) is:

$$(\theta_0)_{\max} = \sin^{-1} (\sqrt{(n_1^2 - n_2^2)/n_0})$$

Hint: For $(\theta_0)_{\max}$, $\phi_i = \phi_c = \sin^{-1} \frac{n_2}{n_1}$, and

$$\sin (\theta_0)_{\max} = \frac{n_1}{n_0} \sin (90 - \phi_c) = \frac{n_1}{n_0} \cdot \cos \phi_c$$

APPENDICES

APPENDIX A 1 THE GREEK ALPHABET

Alpha	A	α	Iota	I	ι	Rho	P	ρ
Beta	B	β	Kappa	K	κ	Sigma	Σ	σ
Gamma	Γ	γ	Lambda	Λ	λ	Tau	T	τ
Delta	Δ	δ	Mu	M	μ	Upsilon	Y	υ
Epsilon	E	ε	Nu	N	ν	Phi	Φ	ϕ, φ
Zeta	Z	ζ	Xi	Ξ	ξ	Chi	X	χ
Eta	H	η	Omicron	O	o	Psi	Ψ	ψ
Theta	Θ	θ	Pi	Π	π	Omega	Ω	ω

APPENDIX A 2 COMMON SI PREFIXES AND SYMBOLS FOR MULTIPLES AND SUB-MULTIPLES

Multiple			Sub-Multiple		
Factor	Prefix	Symbol	Factor	Prefix	symbol
10^{18}	Exa	E	10^{-18}	atto	a
10^{15}	Peta	P	10^{-15}	femto	f
10^{12}	Tera	T	10^{-12}	pico	p
10^9	Giga	G	10^{-9}	nano	n
10^6	Mega	M	10^{-6}	micro	μ
10^3	kilo	k	10^{-3}	milli	m
10^2	Hecto	h	10^{-2}	centi	c
10^1	Deca	da	10^{-1}	deci	d

APPENDIX A 3

SOME IMPORTANT CONSTANTS

Name	Symbol	Value
Speed of light in vacuum	c	$2.9979 \times 10^8 \text{ m s}^{-1}$
Charge of electron	e	$1.602 \times 10^{-19} \text{ C}$
Gravitational constant	G	$6.673 \times 10^{-11} \text{ N m}^2 \text{ kg}^{-2}$
Planck constant	h	$6.626 \times 10^{-34} \text{ J s}$
Boltzmann constant	k	$1.381 \times 10^{-23} \text{ J K}^{-1}$
Avogadro number	N_A	$6.022 \times 10^{23} \text{ mol}^{-1}$
Universal gas constant	R	$8.314 \text{ J mol}^{-1} \text{ K}^{-1}$
Mass of electron	m_e	$9.110 \times 10^{-31} \text{ kg}$
Mass of neutron	m_n	$1.675 \times 10^{-27} \text{ kg}$
Mass of proton	m_p	$1.673 \times 10^{-27} \text{ kg}$
Electron-charge to mass ratio	e/m_e	$1.759 \times 10^{11} \text{ C/kg}$
Faraday constant	F	$9.648 \times 10^4 \text{ C/mol}$
Rydberg constant	R	$1.097 \times 10^7 \text{ m}^{-1}$
Bohr radius	a_0	$5.292 \times 10^{-11} \text{ m}$
Stefan-Boltzmann constant	σ	$5.670 \times 10^{-8} \text{ W m}^{-2} \text{ K}^{-4}$
Wien's Constant	b	$2.898 \times 10^{-3} \text{ m K}$
Permittivity of free space	ϵ_0	$8.854 \times 10^{-12} \text{ C}^2 \text{ N}^{-1} \text{ m}^{-2}$
	$1/4\pi \epsilon_0$	$8.987 \times 10^9 \text{ N m}^2 \text{ C}^{-2}$
Permeability of free space	μ_0	$4\pi \times 10^{-7} \text{ T m A}^{-1}$ $\cong 1.257 \times 10^{-6} \text{ Wb A}^{-1} \text{ m}^{-1}$

OTHER USEFUL CONSTANTS

Name	Symbol	Value
Mechanical equivalent of heat	J	4.186 J cal^{-1}
Standard atmospheric pressure	1 atm	$1.013 \times 10^5 \text{ Pa}$
Absolute zero	0 K	$-273.15 \text{ }^\circ\text{C}$
Electron volt	1 eV	$1.602 \times 10^{-19} \text{ J}$
Unified Atomic mass unit	1 u	$1.661 \times 10^{-27} \text{ kg}$
Electron rest energy	$m_e c^2$	0.511 MeV
Energy equivalent of 1 u	$M c^2$	931.5 MeV
Volume of ideal gas (0 °C and 1 atm)	V	22.4 L mol^{-1}
Acceleration due to gravity (sea level, at equator)	g	9.78049 m s^{-2}

ANSWERS

Chapter 1

- 1.1** $9 \times 10^9 \text{ N}$
- 1.2** $6 \times 10^{-3} \text{ N}$ (repulsive)
- 1.3** (a) 12 cm
(b) 0.2 N (attractive)
- 1.4** 2.4×10^{39} . This is the ratio of electric force to the gravitational force (at the same distance) between an electron and a proton.
- 1.6** Charge is not created or destroyed. It is merely transferred from one body to another.
- 1.8** Zero N
- 1.9** Varies as $1/r^2$.
- 1.11** (a) $5.4 \times 10^6 \text{ N C}^{-1}$ along OB
(b) $8.1 \times 10^{-3} \text{ N}$ along OA
- 1.13** Total charge is zero. Dipole moment = $7.5 \times 10^{-8} \text{ Cm}$ along z-axis.
- 1.14** 10^{-4} N m
- 1.15** (a) 2×10^{12} , from wool to polythene.
(b) Yes, but of a negligible amount ($= 2 \times 10^{-18} \text{ kg}$ in the example).
- 1.16** (a) $1.5 \times 10^{-2} \text{ N}$
(b) 0.24 N
- 1.17** $5.7 \times 10^{-3} \text{ N}$
- 1.18** Charges 1 and 2 are negative, charge 3 is positive. Particle 3 has the highest charge to mass ratio.
- 1.19** $25.98 \text{ N m}^2/\text{C}$
- 1.20** Zero. The number of lines entering the cube is the same as the number of lines leaving the cube.
- 1.21** (a) $0.07 \text{ } \mu\text{C}$
(b) No, only that the net charge inside is zero.
- 1.22** $2.2 \times 10^5 \text{ N m}^2/\text{C}$

1.23 $1.9 \times 10^5 \text{ N m}^2/\text{C}$

1.24 (a) $-10^3 \text{ N m}^2/\text{C}$; because the charge enclosed is the same in the two cases.

(b) -8.8 nC

1.25 -6.67 nC

1.26 (a) $1.45 \times 10^{-3} \text{ C}$

(b) $1.6 \times 10^8 \text{ Nm}^2/\text{C}$

1.27 $10 \text{ } \mu\text{C}/\text{m}$

1.28 (a) Zero, (b) Zero, (c) 1.9 N/C

1.29 (a) Little change in the distribution of charges.

(b) Redistribution of charge on each sphere; positive and negative charges will 'face' each other.

(c) Charge on each sphere will be uniformly distributed.

1.30 $9.81 \times 10^{-4} \text{ mm}$.

1.31 Only (c) is right; the rest cannot represent electrostatic field lines, (a) is wrong because field lines must be normal to a conductor, (b) is wrong because field lines cannot start from a negative charge, (d) is wrong because field lines cannot intersect each other, (e) is wrong because electrostatic field lines cannot form closed loops.

1.32 The force is 10^{-2} N in the negative z -direction, that is in the direction of decreasing electric field. You can check that this is also the direction of decreasing potential energy of the dipole; torque is zero.

1.33 (a) Hint: Choose a Gaussian surface lying wholly within the conductor and enclosing the cavity.

(b) Gauss's law on the same surface as in (a) shows that q must induce $-q$ on the inner surface of the conductor.

(c) Enclose the instrument fully by a metallic surface.

1.34 Hint: Consider the conductor with the hole filled up. Then the field just outside is $(\sigma/\epsilon_0) \hat{n}$ and is zero inside. View this field as a superposition of the field due to the filled up hole plus the field due to the rest of the charged conductor. Inside the conductor, these fields are equal and opposite. Outside they are equal both in magnitude and direction. Hence the field due to the

rest of the conductor is $\left(\frac{\sigma}{2\epsilon_0}\right) \hat{n}$.

1.36 p;uud; n;udd.

1.37 (a) Hint: Prove it by contradiction. Suppose the equilibrium is stable; then the test charge displaced slightly in any direction will experience a restoring force towards the null-point. That is, all field lines near the null point should be directed inwards towards the null-point. That is, there is a net inward flux of electric field through a closed surface around the null-point. But by Gauss's law, the flux of electric field through a surface, not enclosing any charge, must be zero. Hence, the equilibrium cannot be stable.

(b) The mid-point of the line joining the two charges is a null-point. Displace a test charge from the null-point slightly along the line. There is a restoring force. But displace it, say, normal to the line. You will see that the net force takes it away from the null-point. Remember, stability of equilibrium needs restoring force in all directions.

Chapter 2

- 2.1** 10 cm, 40 cm away from the positive charge on the side of the negative charge.
- 2.2** 2.7×10^6 V
- 2.3** (a) 4×10^4 V
(b) 8×10^{-5} J; No
- 2.4** (a) -0.7 J
(b) 0.7 J
- 2.5** (a) The plane normal to AB and passing through its mid-point has zero potential everywhere.
(b) Normal to the plane in the direction AB.
- 2.6** (a) Zero
(b) 10^5 N C $^{-1}$
(c) 4.4×10^4 N C $^{-1}$
- 2.7** 96 pF
- 2.8** (a) 3 pF
(b) 40 V
- 2.9** (a) 9 pF
(b) 2×10^{-10} C, 3×10^{-10} C, 4×10^{-10} C
- 2.10** 18 pF, 1.8×10^{-9} C.
- 2.11** (a) $V = 100$ V, $C = 108$ pF, $Q = 1.08 \times 10^{-8}$ C.
(b) $Q = 1.8 \times 10^{-9}$ C, $C = 108$ pF, $V = 16.6$ V
- 2.12** 1.5×10^{-8} J
- 2.13** 6×10^{-6} J.
- 2.14** 1.2 J; the point R is irrelevant to the answer.
- 2.15** Potential = $4q/\sqrt{3} \pi \epsilon_0 b$; field is zero, as expected by symmetry.
- 2.16** (a) 2.4×10^5 V; 4.0×10^5 Vm $^{-1}$ from charge $2.5 \mu\text{C}$ to $1.5 \mu\text{C}$.
(b) 2.0×10^5 V; 6.6×10^5 Vm $^{-1}$ in the direction that makes an angle of about 69° to the line joining charge $2.5 \mu\text{C}$ to $1.5 \mu\text{C}$.
- 2.17** (a) $-q/(4\pi\epsilon_0 r_1^2)$, $(Q+q)/(4\pi\epsilon_0 r_2^2)$
(b) By Gauss's law, the net charge on the inner surface enclosing the cavity (not having any charge) must be zero. For a cavity of arbitrary shape, this is not enough to claim that the electric field inside must be zero. The cavity may have positive and negative charges with total charge zero. To dispose of this possibility, take a closed loop, part of which is inside the cavity along a field line and the rest inside the conductor. Since field inside the conductor is zero, this gives a net work done by the field in carrying a test charge over a closed loop. We know this is impossible for an electrostatic field. Hence there are no field lines inside the cavity (i.e., no field), and no charge on the inner surface of the conductor, whatever be its shape.
- 2.19** $\lambda/(2\pi\epsilon_0 r)$, where r is the distance of the point from the common axis of the cylinders. The field is radial, perpendicular to the axis.

- 2.20** (a) -27.2 eV .
 (b) 13.6 eV .
 (c) -13.6 eV , 13.6 eV . Note in the latter choice the total energy of the hydrogen atom is zero.
- 2.21** -19.2 eV ; the zero of potential energy is taken to be at infinity.
- 2.22** The ratio of electric field of the first to the second is (b/a) . A flat portion may be equated to a spherical surface of large radius, and a pointed portion to one of small radius.
- 2.23** (a) The atoms in the paper get polarised by the charged comb, resulting in a net force of attraction.
 (b) To enable them to conduct a charge (produced by friction) to the ground.
 (c) Reason similar to (b).
 (d) Current passes only when there is a difference in potential.
- 2.24** (a) On the axis of the dipole, potential is $(\pm 1/4 \pi \epsilon_0) p/(x^2 - a^2)$ where $p = 2qa$ is the magnitude of the dipole moment; the + sign when the point is closer to q and the - sign when it is closer to $-q$. Normal to the axis, at points $(x, y, 0)$, potential is zero.
 (b) The dependence on r is $1/r^2$ type.
 (c) Zero; No, because work done by electrostatic field between two points is independent of the path connecting the two points.
- 2.25** (a) $V_p - V_q > 0$; $V_B - V_A > 0$
 (b) $(\text{P.E.})_q - (\text{P.E.})_p > 0$; $(\text{P.E.})_A - (\text{P.E.})_B > 0$
 (c) Negative
 (d) Positive
 (e) Decreases.
- 2.26** 3 J (Hint: the loss in potential energy of the system of dipoles appears as heat.)
- 2.27** For large r , quadrupole potential goes like $1/r^3$, dipole potential goes like $1/r^2$, monopole potential goes like $1/r$.
- 2.28** Eighteen $1 \mu\text{F}$ capacitors arranged in 6 parallel rows, each row consisting of 3 capacitors in series.
- 2.29** 1130 km^2
- 2.30** Equivalent capacitance $= \frac{200}{3} \text{ pF}$;
 $Q_1 = 10^{-8} \text{ C}$, $V_1 = 100 \text{ V}$; $Q_2 = Q_3 = 10^{-8} \text{ C}$;
 $V_2 = V_3 = 50 \text{ V}$
 $Q_4 = 2.55 \times 10^{-8} \text{ C}$, $V_4 = 200 \text{ V}$.
- 2.31** (a) $2.55 \times 10^{-6} \text{ J}$
 (b) $u = 0.113 \text{ J m}^{-3}$, $u = (1/2) \epsilon_0 E^2$
- 2.32** $2.67 \times 10^{-2} \text{ J}$
- 2.33** Hint: Suppose we increase the separation of the plates by Δx . Work done (by external agency) $= F \Delta x$. This goes to increase the potential energy of the capacitor by $u a \Delta x$ where u is energy density. Therefore $F = u a$ which is easily seen to be $(1/2) QE$, using $u = (1/2) \epsilon_0 E^2$. The physical origin of the factor $1/2$ in the force formula lies in the fact that just outside the conductor, field is E , and inside it is zero. So the average value $E/2$ contributes to the force.

- 2.35** (a) $5.5 \times 10^{-9} \text{ F}$
 (b) $4.5 \times 10^2 \text{ V}$
 (c) $1.3 \times 10^{-11} \text{ F}$.
- 2.36** (a) No, because charge distributions on the spheres will not be uniform.
 (b) No.
 (c) Not necessarily. (True only if the field line is a straight line.) The field line gives the direction of acceleration, not that of velocity, in general.
 (d) Zero, no matter what the shape of the complete orbit.
 (e) No, potential is continuous.
 (f) A single conductor is a capacitor with one of the 'plates' at infinity.
 (g) A water molecule has permanent dipole moment. However, detailed explanation of the value of dielectric constant requires microscopic theory and is beyond the scope of the book.
- 2.37** $1.2 \times 10^{-10} \text{ F}$, $2.9 \times 10^4 \text{ V}$.
- 2.38** 19 cm^2
- 2.39** (a) Planes parallel to x - y plane.
 (b) Same as in (a), except that planes differing by a fixed potential get closer as field increases.
 (c) Concentric spheres centred at the origin.
 (d) A periodically varying shape near the grid which gradually reaches the shape of planes parallel to the grid at far distances.
- 2.40** 30 cm
- 2.41** Hint: By Gauss's law, field between the sphere and the shell is determined by q_1 alone. Hence potential difference between the sphere and the shell is independent of q_2 . If q_1 is positive, this potential difference is always positive.
- 2.42** (a) Our body and the ground form an equipotential surface. As we step out into the open, the original equipotential surfaces of open air change, keeping our head and the ground at the same potential.
 (b) Yes. The steady discharging current in the atmosphere charges up the aluminium sheet gradually and raises its voltage to an extent depending on the capacitance of the capacitor (formed by the sheet, slab and the ground).
 (c) The atmosphere is continually being charged by thunderstorms and lightning all over the globe and discharged through regions of ordinary weather. The two opposing currents are, on an average, in equilibrium.
 (d) Light energy involved in lightning; heat and sound energy in the accompanying thunder.

Chapter 3

- 3.1** 30 A
- 3.2** 17Ω , 8.5 V
- 3.3** (a) 6Ω
 (b) 2 V , 4 V , 6 V
- 3.4** (a) $(20/19) \Omega$
 (b) 10 A , 5 A , 4 A ; 19 A
- 3.5** 1027°C

- 3.6** $2.0 \times 10^{-7} \Omega \text{m}$
- 3.7** $0.0039 ^\circ\text{C}^{-1}$
- 3.8** $867 ^\circ\text{C}$
- 3.9** Current in branch AB = $(4/17) \text{ A}$.
In BC = $(6/17) \text{ A}$, in CD = $(-4/17) \text{ A}$.
In AD = $(6/17) \text{ A}$, in BD. = $(-2/17) \text{ A}$, total current = $(10/17) \text{ A}$
- 3.10** (a) $X = 8.2 \Omega$; to minimise resistance of the connection which are not accounted for in the bridge formula.
(b) 60.5 cm from A.
(c) The galvanometer will show no current.
- 3.11** 11.5 V; the series resistor limits the current drawn from the external source. In its absence, the current will be dangerously high.
- 3.12** Resistance in series = 5988 Ω
- 3.13** Shunt resistance = 10 m Ω
- 3.14** 2.25 V
- 3.15** (a) Electric field is established throughout the circuit almost instantly (with the speed of light) causing at every point a *local electron drift*. Establishment of a current does not have to wait for electrons from one end of the conductor travelling to the other end. However, it does take a little while for the current to reach its steady value.
(b) Each 'free' electron does accelerate increasing its drift speed until it collides with a positive ion of the metal. It loses its drift speed after collision but starts to accelerate and increases its drift speed again only to suffer a collision again and so on. On the average, therefore, electrons acquire only a drift speed.
(c) Simple. Because the electron number density is enormous $\sim 10^{29} \text{ m}^{-3}$.
(d) By no means. The drift velocity is superposed over the large random velocities of electrons.
(e) In the absence of electric field, the paths are straight lines; in the presence of electric field, the paths are, in general, curved.
- 3.16** $2.7 \times 10^4 \text{ s}$ (7.5 h)
- 3.17** (a) Electrons.
(b) electrons and positive ions (protons) of hydrogen.
(c) H^+ and SO_4^{2-} .
(d) same as in (c); the direction of current opposite to that during discharging.
(e) electrons and holes (holes are negative charge vacancies which effectively act as positive charge carriers).
(f) electrons
(g) electrons
Electrons move in a direction opposite to that of the (conventional) current; positive ions and holes move in the direction of the current.
- 3.18** Take the radius of the earth = $6.37 \times 10^6 \text{ m}$ and obtain total charge of the globe. Divide it by current to obtain time = 283 s. Still this method gives you only an estimate; it is not strictly correct. Why?
- 3.19** (a) 1.4 A, 11.9 V
(b) 0.005 A; impossible because a starter motor requires large current ($\sim 100 \text{ A}$) for a few seconds.

3.20 (b) The term 'capacity' of a cell refers to the quantity of electrical energy it can supply. This depends on the size of the cell, among other things. A cell marked 3.5 Ah at 1h discharge rate will supply 3.5 A for one hour. However, depletion of energy is faster (than expected proportionately) for higher currents. Thus the cell will not supply 14 A for 15 min, but may supply 1.75 A for more than 2 h.

(c) (i) An accumulator (secondary cell), say, the lead-acid type, with very low internal resistance (ii) a 'dry battery' (primary cell), say, the carbon-zinc cell.

3.21 $3.9 \times 10^2 \Omega$

3.22 The mass (or weight) ratio of copper to aluminium wire is $(1.72/2.63) \times (8.9/2.7) \approx 2.2$. Since aluminium is lighter, it is preferred for long suspensions of cables.

3.23 Ohm's law is valid to a high accuracy; the resistivity of the alloy manganin is nearly independent of temperature.

3.24 Plot the straight line graph: $\varepsilon - V = Ir$ with $\varepsilon = 1.5 \text{ V}$ and $r = 0.5 \Omega$. The intersection of this line with the (non-linear) $V-I$ plot for the conductor gives the required voltage (0.9 V) and current (1.2 A).

3.25 (a) Inside the cell there is no net field; the electrostatic field due to the plates is balanced by a field of non-electrostatic origin. (If this were not so, charge carriers inside the cell would constitute a current even when the circuit is open). Outside the cell, the net field is just the electrostatic field of the plates.

(b) When the circuit is closed from outside, current flows in the direction of electrostatic field outside, and *opposite* to the direction of electrostatic field inside the cell. The latter fact shows that there is a net field inside the cell *opposite* to the electrostatic field.

(c) They are prevented from being neutralised because the electrostatic field is opposed by a field in the source that has a complex non-electrostatic origin. The latter is related to the emf of the source.

(d) Here the situation is different from (b). Now the current inside the secondary cell during charging is in the direction of electrostatic field. The electrostatic field is greater than the non-electrostatic field. The terminal voltage, therefore, is greater than emf of the secondary cell.

(e) Emf does not have simple electrostatic origin. Therefore, the concept of potential may not be strictly applicable. The emf is work done per unit charge.

3.26 (a) Only current (because it is given to be steady!). The rest depends on the area of cross-section inversely.

(b) No, examples of non-ohmic elements: vacuum diode, semiconductor diode.

(c) Typical electrical insulators (e.g., glass) differ in their resistivity from metals enormously by a factor of the order of 10^{22} . The corresponding factor for thermal insulators versus thermal conductors is only 10^3 .

(d) The internal resistance of a car battery decreases with increase in temperature.

(e) Mainly in its internal resistance, but partly in its emf also.

(f) Because the maximum current drawn from a source $= \varepsilon/r$.

(g) Because, if the circuit is shorted (accidentally), the current drawn will exceed safety limits, if internal resistance is not large.

- 3.27** (a) greater, (b) lower, (c) reduces, (d) decreases
(e) decreases, (f) nearly independent of, (g) 10^{22} .
- 3.28** 1.2Ω
- 3.29** (a) 0.9 A
(b) No, there is no formula for emf and internal resistance of non-similar cells joined in parallel. For this situation, one must use Kirchhoff's rules.
- 3.30** (a) (i) in series, (ii) all in parallel; n^2
(b) (i) Join 1Ω , 2Ω in parallel and the combination in series with 3Ω
(ii) parallel combination of 2Ω and 3Ω in series with 1Ω , (iii) all in series, (iv) all in parallel.
(c) (i) $(16/3) \Omega$, (ii) 5 R .
- 3.31** Hint: Let X be the equivalent resistance of the infinite network. Clearly, $2 + X/(X+1) = X$ which gives $X = (1 + \sqrt{3}) \Omega$; therefore the current is 3.7 A .
- 3.32** (a) Shunt resistance = $4.00 \times 10^{-3} \Omega$ which is also nearly its net resistance; reads slightly less.
(b) Series resistance = 3988Ω ; net resistance = 4000Ω ; reads slightly less.
- 3.33** For X , use (b); for Y , use (a).
- 3.34** (a) The $20 \text{ k}\Omega$ voltmeter gives the highest reading; the last case (both the voltmeters connected across AB) gives the lowest reading.
(b) All cases will give the same reading if internal resistance is zero; otherwise answers similar to (a).
- 3.35** 'Resistance per volt' is a different way of specifying the current at full scale deflection. Here it is $(1/5000) \text{ A} = 0.20 \text{ mA}$. To convert it to a 20 V metre, put a resistance R in series with the metre so that $R \times 2 \times 10^{-4} = 15$ i.e., $R = 7.5 \times 10^4 = 75,000 \Omega$. The 'resistance per volt' of the new metre is the same as before. The higher the 'resistance per volt' of the metre, the less is the current it draws from the circuit and the better it is. So this metre is more accurate than the one graded as $2000 \Omega/\text{V}$.
- 3.36** (a) $\varepsilon = 1.25 \text{ V}$.
(b) To reduce current through the galvanometer when the movable contact is far from the balance point.
(c) No.
(d) No.
(e) No. If ε is greater than the emf of the driver cell of the potentiometer, there will be no balance point on the wire AB.
(f) The circuit, as it is, would be unsuitable, because the balance point (for ε of the order of a few mV) will be very close to the end A and the percentage error in measurement will be very large. The circuit is modified by putting a suitable resistor R in series with the wire AB so that potential drop across AB is only slightly greater than the emf to be measured. Then the balance point will be at larger length of the wire and the percentage error will be much smaller.
- 3.37** $X = 11.75 \Omega$ or 11.8Ω . If there is no balance point, it means potential drops across R or X are greater than the potential drop across the potentiometer wire AB. The obvious thing to do is to reduce current in the outside circuit (and hence potential drops across R and X) suitably by putting a series resistor.
- 3.38** 1.7Ω

- 3.39** (a) Use a standard cell to first calibrate potential drop per unit cm of the potentiometer wire. Then use an external circuit containing a cell and a variable resistor across which the voltmeter to be calibrated is connected. Tap off different potential drops using the variable resistor, measure them by the potentiometer and thus calibrate the voltmeter.
- (b) Here the external circuit will consist of a cell, and a variable resistor in series with a standard resistor and the ammeter. The potential drop across the standard resistor is varied and measured using the calibrated potentiometer. From the known value of the resistance, one obtains different values of the current and thus, the ammeter is calibrated.
- 3.40** 290 k Ω . (The ammeter's reading will be too small). This is an unusual use of a voltmeter meant only for high resistance measurement.
- 3.41** (a) $I = dQ/dt = A d\sigma/dt = \epsilon_0 A dE/dt$
- (b) No; there is no conduction current across the plates. However, if we agree to call $\epsilon_0 A dE/dt$ as 'current' across the capacitor, then the first rule is valid. Maxwell called this the **displacement current**.
- 3.42** (a) About 10 k Ω
- (b) Because our body is sensitive to minute currents even as low as a few mA.
- (c) This is a misleading notion. There is no special attractive force that keeps a person 'stuck' with a high power line. What happens is that a current of the order of 0.05 A or even much less is enough to disorganise our nervous system. The result is that the affected person may lose temporarily his ability to exercise his nervous control to get himself 'free' from the high voltage point.
- (d) The cause of death is not heating, though a person may receive burns if the currents are too large. The cause of death is the interference caused by external currents in our highly sensitive nervous system which is basically electrical in nature. External currents cause convulsive actions, and especially interfere with the nerve processes related to our heartbeating. Beyond a certain point, this interference is fatal.
- (e) About 0.1 V.

Chapter 4

- 4.1** 192 W
- 4.2** 25%
- 4.3** 5 A
- 4.4** 3 A; 70 Ω
- 4.5** 7200 J, Chemical energy stored in the battery.
- 4.6** 2.9 Ω (Hint: Power dissipated as heat is $0.7 \times 50 \times 12 = 420$ W. Therefore, $I^2 R = 420$ W, i.e., $R = 420/144 = 2.9 \Omega$.)
- 4.7** 25 W; 30 kJ
- 4.8** No, because the power dissipated would be 4.5 kW which exceeds the maximum power rating of the resistor.
- 4.9** First heater
- 4.10** Mode (i) involves lesser power wastage, because for the same power transmitted, higher voltage corresponds to smaller current. Therefore lesser power is dissipated as heat in the cables.

- 4.11** (a) From copper to bismuth
(b) From copper to iron
(c) From lead to platinum.
- 4.12** (a) 3.0 W
(b) 0.40 W
(c) 2.6 W
(d) 2.6 W
- 4.13** Series resistor of 8.0 W.
(a) 800 W
(b) 704 W
(c) 86400 J
- 4.14** (a) Power Output = $I^2R = \varepsilon^2R/(R + r)^2$. This is maximum at $R = r$, maximum power output = $\varepsilon^2/4r$.
(b) When the battery is shorted, power output = 0. In this case, the entire power is dissipated as heat inside the battery. Its magnitude is ε^2/r .
- 4.15** 3.0 A
- 4.16** (a) Power output of the source = $\varepsilon I - I^2r$, which is maximum at $I = V/2r$.
(b) $I = (\varepsilon - \varepsilon')/r$, since external resistance R is negligible (ε' stands for back emf). Power output of the electric motor equals power output of the source since $R \approx 0$. The latter is maximum for $I = \varepsilon/2R$. This gives $\varepsilon' = \varepsilon/2$.
- 4.17** (a) 252000 J
(b) 201600 J (Hint: Obtain first the power dissipated as heat. The remaining power is the sum of the mechanical power yielded by the motor and the chemical power stored in the battery.)
- 4.18** Nichrome ribbon, because it has lesser resistance.
- 4.19** Hint: The temperature of the wire increases up to θ , where the heat produced per second by the current equals heat lost (by radiation) per second, i.e.,

$$\frac{I^2 \rho l}{\pi r^2} = h \times 2\pi r l.$$
 (ignoring heat loss from the ends) where h is the heat lost per second per unit surface area of the wire. Clearly, h (and therefore, the steady state temperature θ) is independent of l .
- 4.20** $r = 0.24$ mm. (Hint: From the equation $I^2 \propto r^3$).
- 4.21** (a) No, the steady temperature acquired by a resistor depends not only on the power consumed, but also on its characteristics (such as surface area, emissivity etc.) which determine its power loss due to radiation.
(b) Tungsten bulb.
- 4.22** 120 °C.
- 4.23** 0.25 Ω .
- 4.24** No, the phenomenon described is known as the Peltier effect. In ordinary Joule effect of current (for which the given formula applies), heat is produced by the current no matter what its direction is. (Notice, the formula involves I^2 , i.e., it depends only on the magnitude of I).

Chapter 5

- 5.2 $\pi \times 10^{-4} \text{ T} \approx 3.1 \times 10^{-4} \text{ T}$
- 5.4 $3.5 \times 10^{-5} \text{ T}$
- 5.5 $4 \times 10^{-6} \text{ T}$, vertical up
- 5.6 $1.2 \times 10^{-5} \text{ T}$, towards south
- 5.7 (a) $1.9 \times 10^{-4} \text{ T}$ normal to the plane of the paper going out of it
(b) same magnitude of **B** but opposite in direction to that in (a).
- 5.8 0.6 N m^{-1}
- 5.9 $8.1 \times 10^{-2} \text{ N}$; direction of force given by Fleming's left-hand rule
- 5.10 $2 \times 10^{-5} \text{ N}$; attractive force normal to A towards B
- 5.11 $8\pi \times 10^{-3} \text{ T} \approx 2.5 \times 10^{-2} \text{ T}$
- *5.12 0.96 N m
- 5.13 (a) 1.4 (b) 1
- 5.14 4.2 cm
- 5.15 18 MHz
- 5.16 (a) 3.1 Nm (b) No, the answer is unchanged because the formula $\tau = N I \mathbf{A} \times \mathbf{B}$ is true for a planar loop of any shape.
- 5.17 $5\pi \times 10^{-4} \text{ T} = 1.6 \times 10^{-3} \text{ T}$ towards west
- 5.18 Length about 50 cm, radius about 4 cm, number of turns about 400, current about 10 A. These particulars are not unique. Some adjustment with limits is possible.
- 5.19 (b) In a small region of length $2d$ about the mid-point between the coils,

$$\begin{aligned}
 B &= \frac{\mu_0 I R^2 N}{2} \times \left[\left\{ \left(\frac{R}{2} + d \right)^2 + R^2 \right\}^{-3/2} + \left\{ \left(\frac{R}{2} - d \right)^2 + R^2 \right\}^{-3/2} \right] \\
 &\approx \frac{\mu_0 I R^2 N}{2} \times \left(\frac{5R^2}{4} \right)^{-3/2} \times \left[\left(1 + \frac{4d}{5R} \right)^{-3/2} + \left(1 - \frac{4d}{5R} \right)^{-3/2} \right] \\
 &\approx \frac{\mu_0 I R^2 N}{2R^3} \times \left(\frac{4}{5} \right)^{3/2} \times \left[1 - \frac{6d}{5R} + 1 + \frac{6d}{5R} \right]
 \end{aligned}$$

where in the second and third steps above, terms containing d^2/R^2 and higher powers of d/R are neglected since $\frac{d}{R} \ll 1$. The terms linear in d/R cancel giving a uniform field B in a small region:

$$B = \left(\frac{4}{5} \right)^{3/2} \frac{\mu_0 I N}{R} \approx 0.72 \frac{\mu_0 I N}{R}$$

- 5.20 Hint: B for a toroid is given by the same formula as for a solenoid: $B = \mu_0 n I$, where n in this case is given by $n = \frac{N}{2\pi r}$. The field is non-zero only inside

the core surrounded by the windings. (a) zero (b) 3.0×10^{-2} T (c) zero. Note, the field varies slightly across the cross-section of the toroid as r varies from the inner to outer radius. Answer (b) corresponds to the mean radius $r = 25.5$ cm.

- 5.21** (a) initial \mathbf{v} is either parallel or anti-parallel to \mathbf{B}
 (b) Yes, because magnetic force can change the direction of \mathbf{v} , not its magnitude.
 (c) \mathbf{B} should be in a vertically downward direction.
- 5.22** (a) Electron-positron pair production by a high energy gamma ray
 (b) The charged particles ionise much more in a liquid than in a gas. They progressively lose energy and get slower after collisions with the environment. Since the radius of a track in a given \mathbf{B} is proportional to the speed of the particle, it decreases progressively, giving rise to spiral paths.
- 5.23** (a) Circular trajectory of radius 1.0 mm normal to \mathbf{B}
 (b) helical trajectory of radius 0.5 mm with velocity component 2.3×10^7 m s $^{-1}$ along \mathbf{B} .
- 5.24** $B = 0.66$ T, Kinetic energy = 7.4 MeV.
- 5.25** Deuterium ions or deuterons; the answer is not unique because only the ratio of charge to mass is determined. Other possible answers are He^{++} , Li^{+++} etc.
- 5.26** (a) A horizontal magnetic field of magnitude 0.26 T normal to the conductor in such a direction that Fleming's left-hand rule gives a magnetic force upward.
 (b) 1.176 N.
- 5.27** 1.2 N m $^{-1}$; repulsive. Note, obtaining total force on the wire as $1.2 \times 0.7 = 0.84$ N, is only approximately correct because the formula $F = \frac{\mu_0}{2\pi r} I_1 I_2$ for force per unit length is strictly valid for infinitely long conductors.
- 5.28** (a) 2.1 N vertically downwards
 (b) 2.1 N vertically downwards (true for any angle between current and direction and \mathbf{B} since $l \sin \theta$ remains fixed, equal to 20 cm)
 (c) 1.68 N vertically downwards.
- 5.29** Use $\boldsymbol{\tau} = I\mathbf{A} \times \mathbf{B}$ and $\mathbf{F} = I\mathbf{l} \times \mathbf{B}$
 (a) 1.8×10^{-2} N m along $-y$ direction
 (b) same as in (a)
 (c) 1.8×10^{-2} N m along $-x$ direction
 (d) 1.8×10^{-2} N m at an angle of 240° with the $+x$ direction
 (e) zero
 (f) zero
 Force is zero in each case. Case (e) corresponds to stable, and case (f) corresponds to unstable equilibrium.
- 5.30** (a) zero (b) zero (c) force on each electron is $e\mathbf{vB} = I\mathbf{B}/(nA) = 5 \times 10^{-25}$ N. Note: Answer (c) denotes only the magnetic force.
- 5.31** 108 A
- 5.32** (a) (i) straight line, (ii) parabola, (iii) parabola
 (b) (i) straight line, (ii) circle in a plane normal to the field, (iii) helix with its axis parallel to \mathbf{B} .

The speed is constant for each case under (b)

(c) (i) straight line (ii) same as case (iii) next, (iii) resolve \mathbf{v} along \mathbf{B} and perpendicular to it; motion is circular, motion perpendicular to \mathbf{B} is circular motion parallel to \mathbf{B} (or \mathbf{E}) is accelerated or retarded due to \mathbf{E} .

5.33 (c) Hint: A straight conductor of finite length cannot by itself form a complete steady current circuit. Additional conductors are necessary to close the circuit. These will spoil the symmetry of the problem. The difficulty disappears if the conductor is infinitely long.

5.34 (a) No, because that would require τ to be in a vertical direction. But $\tau = I \mathbf{A} \times \mathbf{B}$, and since \mathbf{A} of the horizontal loop is in the vertical direction, τ must be in the plane of the loop.

(b) Orientation of stable equilibrium is one where the area vector \mathbf{A} of the loop is in the direction of external magnetic field. In this orientation, the magnetic field produced by the loop is in the same direction as external field, both normal to the plane of the loop, thus giving rise to maximum flux of the total field.

(c) It assumes circular shape with its plane normal to the field to maximize flux, since for a given perimeter, a circle encloses greater area than any other shape.

5.35 7.8×10^{-4} N, towards the long conductor if the current in the closer side of the loop is in the same direction as the current in the long conductor; repulsive otherwise. Note that the forces on the two smaller sides of the loop cancel each other.

Chapter 6

6.1 Hint: Potential energy of the configuration arises due to the potential energy of one dipole (say, B) in the magnetic field due to the other (A). Use the result that

$$\mathbf{B}_A = -\frac{\mu_0}{4\pi} \frac{\mathbf{m}_A}{r^3} \quad (\text{on the normal bisector})$$

$$\mathbf{B}_A = \frac{\mu_0}{4\pi} \frac{2\mathbf{m}_A}{r^3} \quad (\text{on the axis})$$

Equilibrium is stable when \mathbf{m}_B is parallel to \mathbf{B}_A , and unstable when \mathbf{m}_B is anti-parallel to \mathbf{B}_A .

(a) AB_1 and AB_2

(b) (i) AB_3 , AB_6 (stable); (ii) AB_5 , AB_4 (unstable)

(c) AB_6

6.2 (a) In either case, one gets two magnets, each with a north and south pole.

(b) Molten iron is above the Curie temperature (770°C) and is, therefore, not ferromagnetic. An iron bar magnet when melted does not retain its magnetism.

(c) No force if the field is uniform. The iron nail experiences a non-uniform magnetic field due to the bar magnet. The induced magnetic moment in the nail, therefore, experiences both force and torque. The net force is attractive because the induced (say) south pole in the nail is closer to the north pole of the magnet than the induced north pole.

(d) Not necessarily. True only if the source of the field has a net non-zero magnetic moment. This is not so for a toroid or even for a straight infinite conductor.

- (e) Depends on what one means by three poles. Poles must always occur in pairs. But one can think of two bar magnets with (say) their north ends glued together as providing a three-pole field configuration.
- (f) Try to bring different ends of the magnets closer. A repulsive force in some situation establishes that both are magnetised. If it is always attractive, then one of them is not magnetised. To see which one, pick up one, say, A and lower one of its ends; first on one of the ends of the other (say, B), and then on the middle of B. If you notice that in the middle of B, A experiences no force, then B is magnetised. If you do not notice any change from the end to the middle of B, then A is magnetised.

- 6.3**
- (a) Magnetic declination, angle of dip, horizontal component of earth's magnetic field.
 - (b) Greater in Britain (about 70°), because Britain is closer to the magnetic north pole.
 - (c) Field lines of \mathbf{B} due to earth's magnetism would seem to come out of the ground.
 - (d) A compass is free to rotate in a horizontal plane, while the earth's field is exactly vertical at the magnetic poles. So the compass can point in any direction there.
 - (e) Use the formula for field \mathbf{B} on the normal bisector of a dipole of magnetic moment \mathbf{m} .

$$\mathbf{B}_A = -\frac{\mu_0}{4\pi} \frac{\mathbf{m}}{r^3}$$

Take $m = 8 \times 10^{22} \text{ J T}^{-1}$, $r = 6.4 \times 10^6 \text{ m}$; one gets $B = 0.3 \text{ G}$, which checks with the order of magnitude of the observed field on the earth.

- (f) Why not? The earth's field is only approximately a dipole field. Local N-S poles may arise due to, for instance, magnetised mineral deposits.

- 6.4**
- (a) Yes, it does change with time. Time scale for appreciable change is roughly a few hundred years. But even on a much smaller scale of a few years, its variations are not completely negligible.
 - (b) Because molten iron (which is the phase of the iron at the high temperatures of the core) is not ferromagnetic.
 - (c) One possibility is the radioactivity in the interior of the earth. But nobody really knows. You should consult a good modern text on geomagnetism for a proper view of the question.
 - (d) Earth's magnetic field gets weakly 'recorded' in certain rocks during solidification. Analysis of this rock magnetism offers clues to geomagnetic history.
 - (e) At large distances, the field gets modified due to the field of ions in motion (in the earth's ionosphere). The latter is sensitive to extra-terrestrial disturbances such as, the solar wind.

- (f) From the relation $R = \frac{mv}{eB}$, an extremely minute field bends charged

particles in a circle of very large radius. Over a small distance, the deflection due to the circular orbit of such large R may not be noticeable, but over the gigantic interstellar distances, the deflection can significantly affect the passage of charged particles, e.g., cosmic rays.

- 6.5**
- (a) Wrong. Magnetic field lines can never emanate from a point as shown. Over any closed surface, the net flux of \mathbf{B} must always be zero; i.e., pictorially as many field lines should seem to enter the surface as the number of lines leaving it and going out. The field lines shown, in fact,

represent electric field of a long positively charged wire. The correct magnetic field lines are lines circling the straight conductor, as described in Chapter 5.

- (b) *Wrong.* Magnetic field lines (like electric field lines) can never cross each other, because otherwise the direction of the field at the point of intersection is ambiguous. There is further error in the figure. Magnetostatic field lines can never form closed loops around empty space. A closed loop of a static magnetic field line must enclose a region across which a current is passing. (By contrast, electrostatic field lines can never form closed loops, neither in empty space, nor even when the loop encloses charges).
- (c) *Right.* Magnetic lines are completely confined within a toroid. Nothing wrong here in field lines forming closed loops, since each loop encloses a region across which a current passes. Note, for clarity of figure, only a few field lines within the toroid have been shown. Actually, the entire region enclosed by the windings contains magnetic field.
- (d) *Wrong.* Field lines due to a solenoid at its ends and outside cannot be so completely straight and confined; such a thing violates Ampere's law. The lines should curve out at the ends, and meet eventually to form closed loops.
- (e) *Right.* These are field lines outside and inside of a bar magnet. Note carefully the direction of field lines inside. Not all field lines emanate out of a north pole (or converge into a south pole). Around both the N-pole, and the S-pole, the net flux of the field is zero.
- (f) *Wrong.* These field lines cannot possibly represent a magnetic field. Look at the upper region. All the field lines seem to emanate out of the shaded plate. The net flux of field through a surface surrounding the shaded plate is not zero. This is impossible for a magnetic field. The given field lines, in fact, show the electrostatic field lines around a positively charged upper plate and a negatively charged lower plate. The difference between figures (e) and (f) should be carefully grasped.
- (g) *Wrong.* Magnetic field lines between two pole pieces cannot be precisely straight at the ends. Some fringing of the lines is inevitable. Otherwise, Ampere's law is violated. This is also true for electric field lines.

6.6 (a) No. The magnetic force is always normal to \mathbf{B} (remember magnetic force $= e\mathbf{v} \times \mathbf{B}$). It is misleading to call field lines of \mathbf{B} as lines of force.

- (b) If field lines were entirely confined between two ends of a straight solenoid, the flux through the cross-section at each end would be non-zero. But the flux of field \mathbf{B} through any closed surface must always be zero. For a toroid, this difficulty is absent because it has no 'ends'.
- (c) Gauss's law of magnetism states that the flux of \mathbf{B} through any closed

surface is always zero: $\oint_S \mathbf{B} \cdot d\mathbf{s} = 0$ for any closed S .

If monopoles existed, the r.h.s. would equal the monopole (magnetic charge) q_m enclosed by S . (analogous to Gauss's law of

electrostatics) $\oint_S \mathbf{B} \cdot d\mathbf{s} = \frac{q_m}{\mu_0}$, where q_m is the (monopole) magnetic charge included by S .

- (d) No. There is no force or torque on an element due to the field produced by that element itself. But there is a force (or torque) on an element of the same wire. (For the special case of a straight wire, this force is zero).

- (e) Yes. The average of the charges in the system may be zero. Yet, the mean of the magnetic moments due to various current loops may not be zero. A neutron, for example, has zero charge but non-zero magnetic moment.
- (f) This is not a simple question. Think of the nail as so many charges in motion. In the bar's magnetic field, each charge in motion in the nail experiences a magnetic force that alters its velocity but not speed. The total energy the system (the nail) cannot change, since magnetic force does no work. But because of changes in individual velocity directions, the velocity of the centre of mass can increase, obviously at the expense of the nail's internal energy. Magnetic field provides force, while internal energy of the nail provides increase in kinetic energy of the nail as a whole. The above answer would be complete if magnetic moments arose only from the (orbital) motion of charges inside the nail. Actually, however, charges (electrons) have intrinsic magnetic moments also — a fact that can be understood only in the framework of quantum mechanics. In a non-uniform field, these magnetic dipoles experience force leading to an increase in the kinetic energy of the nail as a whole.

6.7 (a) The tendency to disrupt the alignment of dipoles (with the magnetising field) arising from random thermal motion is reduced at lower temperatures.

(b) The induced dipole moment in a diamagnetic sample is always opposite to the magnetising field, no matter what the internal motion of the atoms is.

(c) Slightly less, since bismuth is diamagnetic.

(d) No, as it evident from the magnetisation curve. From the slope of magnetisation curve, it is clear that μ is greater for lower fields.

(e) Proof of this important fact (of much practical use) is based on boundary conditions of magnetic fields (**B** and **H**) at the interface of two media. (When one of the media has $\mu \gg 1$, the field lines meet this medium nearly normally). Details are beyond the scope of this book.

(f) Yes. Apart from minor differences in strength of the individual atomic dipoles of two different materials, a paramagnetic sample with saturated magnetisation will have the same order of magnetisation. But of course, saturation requires impractically high magnetising fields.

6.8 (b) Carbon steel piece, because heat lost per cycle is proportional to the area of hysteresis loop.

(c) Magnetisation of a ferromagnet is not a single-valued function of the magnetising field. Its value for a particular field depends both on the field and also on history of magnetisation (i.e., how many cycles of magnetisation it has gone through etc.). In other words, the value of magnetisation is a record or 'memory' of its cycles of magnetisation. If information bits can be made to correspond to these cycles, the system displaying such a hysteresis loop can act as a device for storing information.

(d) Ceramics, (specially treated barium iron oxides) also called ferrites.

(e) Surround the region by soft iron rings. Magnetic field lines will be drawn into the rings, and the enclosed space will be free of magnetic field. But this shielding is only approximate, unlike the perfect electric shielding of a cavity in a conductor placed in an external electric field.

6.9 0.36 J T^{-1}

6.10 (a) **m** parallel to **B**; $U = -mB = -4.8 \times 10^{-2} \text{ J}$; stable

(b) **m** anti-parallel to **B**; $U' = +mB = +4.8 \times 10^{-2} \text{ J}$; unstable

- 6.11** 0.60 JT^{-1} along the axis of the solenoid; the direction determined by the sense of flow of current.
- 6.12** $7.5 \times 10^{-2} \text{ J}$
- 6.13** (a) (i) 0.33 J (ii) 0.66 J
 (b) (i) torque of magnitude 0.33 J in a direction that tends to align the magnetic moment vector along **B**. (ii) Zero.
- 6.14** (a) 1.28 A m^2 along the axis in the direction related to the sense of current via the right-handed screw rule.
 (b) force is zero in uniform field; torque = 0.048 Nm in a direction that tends to align the axis of the solenoid (i.e., its magnetic moment vector) along **B**.
- 6.15** $B = 0.35 \text{ sec } 22^\circ \approx 0.38 \text{ G}$.
- 6.16** The earth's field lies in a vertical plane 12° west of the geographic meridian making an angle of 60° (upwards) with the horizontal (magnetic south to magnetic north) direction. Magnitude = 0.32 G .
- 6.17** (i) 0.96 G along S-N direction.
 (ii) 0.48 G along N-S direction.
- 6.18** 0.54 G in the direction of earth's field.
- 6.19** At $14 \times 2^{-1/3} \approx 11.1 \text{ cm}$ on the normal bisector
- 6.20** (a) $(\mu_0 m)/(4 \pi r^3) = 0.42 \times 10^{-4}$ which gives $r = 5.0 \text{ cm}$.
 (b) $(2\mu_0 m)/(4 \pi r_1^3) = 0.42 \times 10^{-4}$ i.e., $r_1 = 2^{1/3} r \approx 6.3 \text{ cm}$.
- 6.21** Use $I = mB/(4 \pi^2 v^2)$; $m = N i A$ to get $I = 1.2 \times 10^{-4} \text{ kg m}^2$
- 6.22** Parallel to and above the cable at a distance of 1.5 cm .
- 6.23** Below the cable:
 $R_h = 0.39 \cos 35^\circ - 0.2$
 $= 0.12 \text{ G}$
 $R_v = 0.39 \sin 35^\circ = 0.22 \text{ G}$
 $R = \sqrt{R_h^2 + R_v^2} = 0.25 \text{ G};$
 $\theta = \tan^{-1} \frac{R_v}{R_h} \approx 62^\circ$
- Above the cable :
 $R_h = 0.39 \cos 35^\circ + 0.2$
 $= 0.52 \text{ G}$
 $R_v = 0.224 \text{ G}$
 $R = 0.57 \text{ G}, \theta \approx 23^\circ$
- 6.24** (a) $B_h = (\mu_0 i N / 2r) \cos 45^\circ = 0.39 \text{ G}$
 (b) east to west (i.e., the needle will reverse its original direction).
- 6.25** Magnitude of the other field

$$= \frac{1.2 \times 10^{-2} \times \sin 15^\circ}{\sin 45^\circ}$$

$$= 4.4 \times 10^{-3} \text{ T}$$

6.26 $R = \frac{meV}{eB}$

$$= \frac{\sqrt{2m_e \times \text{kinetic energy}}}{eB}$$

$$= 11.3 \text{ m}$$

Up or down deflection = $R(1 - \cos \theta)$ where $\sin \theta = 0.3/11.3$. We get deflection $\approx 4 \text{ mm}$.

6.27 Initially, total dipole moment

$$= 0.15 \times 1.5 \times 10^{-23} \times 2.0 \times 10^{24}$$

$$= 4.5 \text{ J T}^{-1}$$

Use Curie's Law $m \propto B/T$ to get the final dipole moment

$$= 4.5 \times (0.98/0.84) \times (4.2/2.8)$$

$$= 7.9 \text{ J T}^{-1}$$

6.28 Use the formula $B = \mu_0 K n i$, $K = \mu/\mu_0$ (relative permeability) to get $B = 4.48 \text{ T}$.

6.29 Of the two, the relation $\mu_l = -(e/2m)\mathbf{l}$ is in accordance with classical physics. It follows easily from the definitions of μ_l and \mathbf{l} :

$$\mu_l = iA = (e/T)\pi r^2$$

$$\mathbf{l} = m\mathbf{v}r = m \frac{2\pi r^2}{T}$$

where r is the radius of the circular orbit which the electron of mass m and charge $(-e)$ completes in time T . Clearly, $\mu_l/\mathbf{l} = e/2m$

Since charge of the electron is negative ($= -e$), it is easily seen that μ_l and \mathbf{l} are antiparallel, both normal to the plane of the orbit. Therefore, $\mu_l = -e/2m \mathbf{l}$. Note μ_s/S in contrast to μ_l/\mathbf{l} is e/m , i.e., twice the classically expected value. This latter result (verified experimentally) is an outstanding consequence of modern quantum theory and cannot be obtained classically.

Chapter 7

7.1 (a) Along abcd [motion causes increase in flux into the loop, so induced current tends to decrease the flux into the loop].

(b) Along acba

(c) Along adcb

(d) Along abcd

Direction of induced current same as expected on magnetic force considerations.

Note: induced current ceases when the loop is completely in or out.

7.2 (a) Along pq

(b) Along qp, along xy

(c) Along xyz

(d) Along zyx

(e) Along xy

(f) No induced current since field lines lie in the plane of the loop.

7.3 (a) Along adcb (flux into increases during shape change, so induced current produces flux out.)

(b) Along adcb (flux out decreases during shape change, so induced current produces flux out.)

- 7.4** (a) No; current is induced only if there is a *change* in the flux linking the loop.
 (b) No current is induced in *either* case. Current is induced due to changing magnetic (and not electric) flux.
 (c) In the rectangular loop. (For the circular loop, the rate of change of area of the loop inside the field region is not constant.)
 (d) a will be positive relative to b.

7.5 $1.6 \times 10^{-3} \text{ V}$

7.6 $7.6 \times 10^{-6} \text{ V}$

7.7 100 V

7.8 Flux through each turn of the loop $= \pi r^2 B \cos(\omega t)$

$$\varepsilon = -N \omega \pi r^2 B \sin(\omega t)$$

$$\varepsilon_{\max} = -N \omega \pi r^2 B$$

$$= 20 \times 50 \times \pi \times 64 \times 10^{-4} \times 3.0 \times 10^{-2} = 0.603 \text{ V}$$

ε_{avg} is zero over a cycle

$$I_{\max} = 0.0603 \text{ A}$$

$$\text{Power loss} = (1/2) \varepsilon_{\max} I_{\max} = 0.036 \text{ W}$$

The induced current causes a torque opposing the rotation of the coil. An external agent (rotor) must supply torque (and do work) to counter this torque in order to keep the coil rotating uniformly. Thus, the source of the power dissipated as heat in the coil is the external rotor.

7.9 Induced emf $= \frac{1}{2} \omega B R^2$

$$= (1/2) \times 4\pi \times 0.4 \times 10^{-4} \times (0.5)^2$$

$$= 6.28 \times 10^{-5} \text{ V.}$$

The number of spokes is immaterial, because the emf's across the spokes are 'in parallel'.

7.10 (1) $2.4 \times 10^{-4} \text{ V}$, lasting 2 s.

(2) $0.6 \times 10^{-4} \text{ V}$, lasting 8 s.

7.11 Induced emf $= 8 \times 2 \times 10^{-4} \times 0.02 = 3.2 \times 10^{-5} \text{ V}$

Induced current $= 2 \times 10^{-5} \text{ A}$

Power loss $= 6.4 \times 10^{-10} \text{ W}$

Source of this power is the external agent responsible for changing the magnetic field with time.

7.12 Rate of change of flux due to explicit time variation in B

$$= 144 \times 10^{-4} \text{ m}^2 \times 10^{-3} \text{ T s}^{-1}$$

$$= 1.44 \times 10^{-5} \text{ Wb s}^{-1}$$

Rate of change of flux due to motion of the loop in a non-uniform B ,

$$= 144 \times 10^{-4} \text{ m}^2 \times 10^{-3} \text{ T cm}^{-1} \times 8 \text{ cm s}^{-1}$$

$$= 11.52 \times 10^{-5} \text{ Wb s}^{-1}$$

The two effects add up since both cause a *decrease* in flux along the positive z -direction. Therefore, induced emf $= 12.96 \times 10^{-5} \text{ V}$; induced current $= 2.88 \times 10^{-2} \text{ A}$. The direction of induced current is such as to *increase* the flux through the loop along positive z -direction. If for the observer the loop moves to the right, the current will be seen to be anti-clockwise. A proper proof of the procedure above is as follows:

$$\Phi(t) = \int_0^a B(x, t) dx$$

$$\frac{d\Phi}{dt} = a \int_0^a dx \frac{dB(x, t)}{dt}$$

using,

$$\begin{aligned} \frac{dB}{dt} &= \frac{\partial B}{\partial t} + \frac{\partial B}{\partial x} \frac{dx}{dt} \\ &= \left[\frac{\partial B}{\partial t} + v \frac{\partial B}{\partial x} \right] \end{aligned}$$

we get,

$$\begin{aligned} \frac{d\Phi}{dt} &= a \int_0^a dx \left[\frac{\partial B(x, t)}{\partial t} + v \frac{\partial B(x, t)}{\partial x} \right] \\ &= A \left[\frac{\partial B}{\partial t} + v \frac{\partial B}{\partial x} \right] \end{aligned}$$

where $A = a^2$

The last step follows because $(\frac{\partial B}{\partial t})$, $(\frac{\partial B}{\partial x})$ and v are given to be constants in the problem. Even if you do not understand this formal proof (which requires good familiarity with calculus), you will still appreciate that flux change can occur both due to the motion of the loop as well as time variations in the magnetic field.

$$\begin{aligned} \text{7.13 } Q &= \int_{t_i}^{t_f} I dt \\ &= \frac{1}{R} \int_{t_i}^{t_f} \mathcal{E} dt \\ &= -\frac{N}{R} \int_{\Phi_i}^{\Phi_f} d\Phi \\ &= \frac{N}{R} (\Phi_i - \Phi_f) \end{aligned}$$

for $N = 25$, $R = 0.50 \, \Omega$, $Q = 7.5 \times 10^{-3} \text{ C}$,

$\Phi_f = 0$, $A = 2.0 \times 10^{-4} \text{ m}^2$, $\Phi_i = 1.5 \times 10^{-4} \text{ Wb}$,

$B = \Phi_i / A = 0.75 \text{ T}$

$$\text{7.14 } |\mathcal{E}| = vBl = 0.12 \times 0.50 \times 0.15 = 9.0 \text{ mV};$$

P positive end and Q negative end.

(b) Yes. When K is closed, the excess charge is maintained by the continuous flow of current.

(c) Magnetic force is cancelled by the electric force set up due to the excess charge of opposite signs at the ends of the rod.

(d) Retarding force $= IBl$

$$= \frac{9 \text{ mV}}{9 \text{ m}\Omega} \times 0.5 \text{ T} \times 0.15 \text{ m}$$

$$= 75 \times 10^{-3} \text{ N.}$$

- (e) Power expended by an external agent against the above retarding force to keep the rod moving uniformly at 12 cm s^{-1}

$$= 75 \times 10^{-3} \times 12 \times 10^{-2} = 9.0 \times 10^{-3} \text{ W}$$

When K is open, no power is expended.

- (f) $I^2 R = 1 \times 1 \times 9 \times 10^{-3} = 9.0 \times 10^{-3} \text{ W}$

The source of this power is the power provided by the external agent as calculated above.

- (g) Zero; motion of the rod does not cut across the field lines. [Note: length of PQ has been considered above to be equal to the spacing between the rails.]

7.15 $B = \frac{\mu_0 n I}{l}$

(Inside the solenoid away from the ends.)

$$\Phi = \frac{\mu_0 n I}{l} A$$

Total flux linkage $= N\Phi$

$$= \frac{\mu_0 n^2 A}{l} I$$

(Ignoring end variations in **B**).

$$|\mathcal{E}| = \frac{d}{dt} (N\Phi),$$

$$|\mathcal{E}|_{\text{av}} = \frac{\text{total change in flux}}{\text{total time}}$$

$$|\mathcal{E}|_{\text{av}} = \frac{4\pi \times 10^{-7} \times 25 \times 10^{-4}}{0.3 \times 10^{-3}} \times (500)^2 \times 2.5$$

$$= 6.5 \text{ V}$$

- 7.16** Vertical component of **B**

$$= 5.0 \times 10^{-4} \sin 30^\circ$$

$$= 2.5 \times 10^{-4} \text{ T}$$

$$l = 25 \text{ m}$$

$$\mathcal{E} = 500 \times 2.5 \times 10^{-4} \times 25$$

$$= 3.1 \text{ V}$$

The direction of the wing is immaterial (as long as it is horizontal) for this answer.

7.17 $M = \frac{\mu_0 a}{2\pi} \ln \left(1 + \frac{x}{a} \right)$

$$\mathcal{E} = 1.7 \times 10^{-5} \text{ V}$$

7.18 $-\frac{B\pi a^2 \lambda}{mR} \hat{\mathbf{k}}$

Chapter 8

8.1 (a) 2.20 A

(b) 484 W

8.2 (a) $\frac{300}{\sqrt{2}} = 212.1 \text{ V}$

(b) $10\sqrt{2} = 14.1 \text{ A}$

8.3 15.9 A

8.4 2.49 A

8.5 Zero in each case.

8.7 125 s^{-1} ; 25

8.8 A choke coil reduces voltage across the tube without wasting power. A resistor would waste power as heat.

8.11 $1.1 \times 10^3 \text{ s}^{-1}$

8.12 0.6 J, same at later times.

8.13 2,000 W

8.14 $v = \frac{1}{2\pi} \sqrt{\frac{1}{LC}}$, i.e., $C = \frac{1}{4\pi^2 v^2 L}$

For $L = 200 \text{ } \mu\text{H}$, $v = 1200 \text{ kHz}$, $C = 87.9 \text{ pF}$.

For $L = 200 \text{ } \mu\text{H}$, $v = 800 \text{ kHz}$, $C = 197.8 \text{ pF}$.

The variable capacitor should have a range of about 88 pF to 198 pF.

8.15 (a) 50 rad s^{-1}

(b) $40 \text{ } \Omega$, 8.1 A

(c) $V_{Lrms} = 1437.5 \text{ V}$, $V_{Crms} = 1437.5 \text{ V}$, $V_{Rrms} = 230 \text{ V}$

$$V_{LCrms} = I_{rms} \left(\omega_0 L - \frac{1}{\omega_0 C} \right) = 0$$

8.16 (a) 1.0 J. Yes, sum of the energies stored in L and C is conserved if $R = 0$.

(b) $\omega = 10^3 \text{ rad s}^{-1}$, $v = 159 \text{ Hz}$

(c) $q = q_0 \cos \omega t$

(i) energy stored is completely electrical at $t = 0, \frac{T}{2}, T, \frac{3T}{2}, \dots$

(ii) energy stored is completely magnetic (i.e., electrical energy is zero)

at $t = \frac{T}{4}, \frac{3T}{4}, \frac{5T}{4}, \dots$, where $T = \frac{1}{v} = 6.3 \text{ ms}$.

(d) At $t = \frac{T}{8}, \frac{3T}{8}, \frac{5T}{8}, \dots$, because $q = q_0 \cos \frac{\omega T}{8} = q_0 \cos \frac{\pi}{4} = \frac{q_0}{\sqrt{2}}$.

Therefore, electrical energy $= \frac{q^2}{2C} = \frac{1}{2} \left(\frac{q_0^2}{2C} \right)$ which is half the total energy.

(e) R damps out the LC oscillations eventually. The whole of the initial energy ($= 1.0 \text{ J}$) is eventually dissipated as heat.

8.17 For an LR circuit, if $V = V_0 \sin \omega t$

$$I = \frac{V_0}{\sqrt{R^2 + \omega^2 L^2}} \sin(\omega t - \phi), \text{ where } \tan \phi = (\omega L / R).$$

(a) $I_0 = 1.82 \text{ A}$

(b) V is maximum at $t = 0$, I is maximum at $t = (\phi / \omega)$.

$$\text{Now, } \tan \phi = \frac{2\pi \nu L}{R} = 1.571 \quad \text{or } \phi \approx 57.5^\circ$$

$$\text{Therefore, time lag} = \left(\frac{57.5\pi}{180} \right) \times \frac{1}{2\pi \times 50} = 3.2 \text{ ms}$$

8.18 (a) $I_0 = 1.1 \times 10^{-2} \text{ A}$

(b) $\tan \phi = 100\pi$, ϕ is close to $\pi/2$.

I_0 is much smaller than the low frequency case (Exercise 8.17) showing thereby that at high frequencies, L nearly amounts to an open circuit. In a dc circuit (after steady state) $\omega = 0$, so here L acts like a pure conductor.

8.19 For a RC circuit, if $V = V_0 \sin \omega t$

$$I = \frac{V_0}{\sqrt{R^2 + (1/\omega C)^2}} \sin(\omega t + \phi) \quad \text{where } \tan \phi = \frac{1}{\omega C R}$$

(a) $I_0 = 3.23 \text{ A}$

(b) $\phi = 33.5^\circ$

$$\text{Time lag} = \frac{\phi}{\omega} = 1.55 \text{ ms}$$

8.20 (a) $I_0 = 3.88 \text{ A}$

(b) $\phi \approx 0.2$ and is nearly zero at high frequency. Thus, at high frequency, C acts like a conductor. For a dc circuit, after steady state, $\omega = 0$ and C amounts to an open circuit.

8.21 Effective impedance of the parallel LCR circuit is given by

$$\frac{1}{Z} = \sqrt{\frac{1}{R^2} + \left(\omega C - \frac{1}{\omega L} \right)^2}$$

$$\text{which is minimum at } \omega = \omega_0 = \frac{1}{\sqrt{LC}}$$

Therefore, $|Z|$ is maximum at $\omega = \omega_0$, and the total current amplitude is minimum.

In R branch, $I_{R_{rms}} = 5.75 \text{ A}$

In L branch, $I_{L_{rms}} = 0.92 \text{ A}$

In C branch, $I_{C_{rms}} = 0.92 \text{ A}$

Note, total current $I_{rms} = 5.75 \text{ A}$, since the currents in L and C branch are 180° out of phase and add to zero at every instant of the cycle.

8.22 (a) For $V = V_0 \sin \omega t$

$$I = \frac{V_0}{\left| \omega L - \frac{1}{\omega C} \right|} \sin \left(\omega t + \frac{\pi}{2} \right); \quad \text{if } R = 0$$

where $-$ sign appears if $\omega L > 1/\omega C$, and $+$ sign appears if $\omega L < 1/\omega C$.
 $I_0 = 11.6 \text{ A}$, $I_{\text{rms}} = 8.24 \text{ A}$

(b) $V_{L\text{rms}} = 207 \text{ V}$, $V_{C\text{rms}} = 437 \text{ V}$

(Note: $437 \text{ V} - 207 \text{ V} = 230 \text{ V}$ is equal to the applied rms voltage as should be the case. The voltage across L and C gets subtracted because they are 180° out of phase.)

- (c) Whatever be the current I in L , actual voltage leads current by $\pi/2$. Therefore, average power consumed by L is zero.
 (d) For C , voltage lags by $\pi/2$. Again, average power consumed by C is zero.
 (e) Total average power absorbed is zero.

8.23 $I_{\text{rms}} = 7.26 \text{ A}$

Average power to $R = I_{\text{rms}}^2 R = 791 \text{ W}$

Average power to $L =$ Average power to $C = 0$

Total power absorbed = 791 W

8.24 (a) $\omega_0 = 4167 \text{ rad s}^{-1}$; $\nu_0 = 663 \text{ Hz}$

$$I_0^{\text{max}} = 14.1 \text{ A}$$

(b) $\bar{P} = (1/2)I_0^2 R$ which is maximum at the same frequency (663 Hz) for which I_0 is maximum $\bar{P}_{\text{max}} = (1/2)(I_{\text{max}})^2 R = 2300 \text{ W}$.

(c) At $\omega = \omega_0 \pm \Delta\omega$ [Approximation good if $(R/2L) \ll \omega_0$].

$$\Delta\omega = R/2L = 95.8 \text{ rad s}^{-1}; \quad \Delta\nu = \Delta\omega/2\pi = 15.2 \text{ Hz}.$$

Power absorbed is half the peak power at $\nu = 648 \text{ Hz}$ and 678 Hz .

At these frequencies, current amplitude is $(1/\sqrt{2})$ times I_0^{max} , i.e., current amplitude (at half the peak power points) is 10 A .

(d) $Q = 21.7$

8.25 $\omega_0 = 111 \text{ rad s}^{-1}$; $Q = 45$

To double Q without changing ω_0 , reduce R to 3.7Ω .

8.26 (a) Yes. The same is *not* true for rms voltage, because voltages across different elements may not be in phase. See, for example, answer to Exercise 8.22.

- (b) To supply a given power, low power-factor means a large current is needed. This causes larger heat losses due to the factor $I^2 R$.
 (c) Power factor = (R/Z) . Many ac machines have inductive reactance. A capacitance of appropriate value reduces the net reactance so that Z approaches R .
 (d) The high induced voltage, when the circuit is broken, is used to charge the capacitor, thus avoiding sparks, etc.
 (e) For dc, impedance of L is negligible and of C very high (infinite), so the dc signal appears across C . For high frequency ac, impedance of L is high and that of C is low. So, the ac signal appears across L .

- (f) For a steady state dc, L has no effect, even if it is increased by an iron core. For ac, the lamp will shine dimly because of additional impedance of the choke. It will dim further when the iron core is inserted which increases the choke's impedance.
- (g) For dc, capacitor is an open circuit. The lamp will not shine at all, even if C is reduced. For ac, the lamp will shine because C 'conducts' ac. Reducing C , will increase impedance of C and the lamp will shine less brightly than before.

8.27 400

8.28 Hydroelectric power = $h\rho g \times A \times v = h\rho g \beta$

where $\beta = Av$ is the flow (volume of water flowing per second across a cross-section).

$$\text{Electric power available} = 0.6 \times 300 \times 10^3 \times 9.8 \times 100 \text{ W} \\ = 176 \text{ MW}$$

8.29 Line resistance = $30 \times 0.5 = 15 \Omega$.

$$\text{rms current in the line} = \frac{800 \times 1000 \text{ W}}{4000 \text{ V}} = 200 \text{ A}$$

(a) Line power loss = $(200 \text{ A})^2 \times 15 \Omega = 600 \text{ kW}$.

(b) Power supply by the plant = $800 \text{ kW} + 600 \text{ kW} = 1400 \text{ kW}$.

(c) Voltage drop on the line = $200 \text{ A} \times 15 \Omega = 3000 \text{ V}$.

The step-up transformer at the plant is $440 \text{ V} - 7000 \text{ V}$.

$$\text{Current} = \frac{800 \times 1000 \text{ W}}{40,000 \text{ V}} = 20 \text{ A}$$

(a) Line power loss = $(20 \text{ A})^2 \times (15 \Omega) = 6 \text{ kW}$.

(b) Power supply by the plant = $800 \text{ kW} + 6 \text{ kW} = 806 \text{ kW}$.

(c) Voltage drop on the line = $20 \text{ A} \times 15 \Omega = 300 \text{ V}$.

The step-up transformer is $440 \text{ V} - 40,300 \text{ V}$. It is clear that percentage power loss is greatly reduced by high voltage transmission. In Exercise 8.29, this power loss is $(600/1400) \times 100 = 43\%$. In this exercise, it is only $(6/806) \times 100 = 0.74\%$.

Chapter 9

9.1 The speed in vacuum is the same for all : $c = 3 \times 10^8 \text{ m s}^{-1}$

9.2 \mathbf{E} and \mathbf{B} in x - y plane and are mutually perpendicular, 10 m.

9.3 Wavelength band : 40 m - 25 m.

9.4 10^9 Hz

9.5 153 N/C.

9.6 (a) 400 nT, $3.14 \times 10^8 \text{ rad/s}$, 1.05 rad/s, 6.00 m.

(b) $\mathbf{E} = \{ (120 \text{ N/C}) \sin[(1.05 \text{ rad/m})x - (3.14 \times 10^8 \text{ rad/s})t] \} \hat{\mathbf{j}}$

$\mathbf{B} = \{ (400 \text{ nT}) \sin[(1.05 \text{ rad/m})x - (3.14 \times 10^8 \text{ rad/s})t] \} \hat{\mathbf{k}}$

9.7 Photon energy (for $\lambda = 1 \text{ m}$)

$$= \frac{6.63 \times 10^{-34} \times 3 \times 10^8}{1.6 \times 10^{-19}} \text{ eV} = 1.24 \times 10^{-6} \text{ eV}$$

Photon energy for other wavelengths in the figure for em spectrum can be obtained by multiplying approximate powers of ten. Energy of a photon that a source produces indicates the spacings of the relevant energy levels of the source. For example, $\lambda = 10^{-12}$ m corresponds to photon energy $= 1.24 \times 10^6$ eV $= 1.24$ MeV. This indicates that nuclear energy levels (transition between which causes γ -ray emission) are typically spaced by 1 MeV or so. Similarly, a visible wavelength $\lambda = 5 \times 10^{-7}$ m, corresponds to photon energy $= 2.5$ eV. This implies that energy levels (transition between which gives visible radiation) are typically spaced by a few eV.

9.8 (a) $\lambda = (c/\nu) = 1.5 \times 10^{-2}$ m

(b) $B_0 = (E_0/c) = 1.6 \times 10^{-7}$ T

(c) Energy density in **E** field : $u_E = (1/2)\epsilon_0 E^2$

Energy density in **B** field : $u_B = (1/2\mu_0)B^2$

Using $E = cB$, and $c = \frac{1}{\sqrt{\mu_0\epsilon_0}}$, $u_E = u_B$

9.9 (a) $-\hat{j}$, (b) 3.5 m, (c) 86 MHz, (d) 100 nT,

(e) $\{(100 \text{ nT}) \cos[(1.8 \text{ rad/m})y + (5.4 \times 10^6 \text{ rad/s})t]\} \hat{k}$

9.10 (a) 0.4 W/m^2 , (b) 0.004 W/m^2

9.11 A body at temperature T produces a continuous spectrum of wavelengths. For a black body, the wavelength corresponding to maximum intensity of radiation is given according to Planck's law by the relation: $\lambda_m = 0.29 \text{ cm K/T}$. For $\lambda_m = 10^{-6}$ m, $T = 2900$ K. Temperatures for other wavelengths can be found. These numbers tell us the temperature ranges required for obtaining radiations in different parts of the em spectrum. Thus, to obtain visible radiation, say $\lambda = 5 \times 10^{-7}$ m, the source should have a temperature of about 6000 K. Note, a lower temperature will also produce this wavelength but not the maximum intensity.

9.12 There is no contradiction. Field lines *inside* the bar magnet go away from S towards N. The net flux of **B** over any surface fully enclosing N or S must be identically zero.

9.13 (a) Radio (short wavelength end)

(b) Radio (short wavelength end)

(c) Microwave

(d) Visible (Yellow)

(e) X-rays (or soft γ -rays) region

9.14 (a) Ionosphere reflects waves in these bands.

(b) Television signals are not properly reflected by the ionosphere (see text). Therefore, reflection is effected by satellites.

(c) Atmosphere absorbs X-rays, while visible and radiowaves can penetrate it.

(d) It absorbs ultraviolet radiations from the sun and prevents it from reaching the earth's surface and causing damage to life.

(e) The temperature of the earth would be lower because the Greenhouse effect of the atmosphere would be absent.

(f) The clouds produced by global nuclear war would perhaps cover substantial parts of the sky preventing solar light from reaching many parts of the globe. This would cause a 'winter'.

Chapter 10

- 10.1** Because of inverse-square dependence on distance, light energy falling per second reduces by a factor of 4. Therefore, to receive the same amount of light, exposure time should be increased to $(2.5 \text{ s}) \times 4 = 10 \text{ s}$.
- 10.2** The screen should be placed 54 cm from the mirror. The image is real, inverted and magnified. The size of the image is 5.0 cm. If the candle is moved closer, the screen would have to be moved farther and farther. Closer than 18 cm from the mirror, the image gets virtual and cannot be collected on the screen.
- 10.3** Virtual image located 6.7 cm behind the mirror. Magnification = $5/9$, i.e., the size of the image is reduced to $(5/9) \times 4.5 \text{ cm} = 2.5 \text{ cm}$. As the needle is moved farther from the mirror, the image moves towards the focus (but never beyond) and gets progressively diminished in size.
- 10.4** Image located at $16 \frac{2}{3} \text{ cm}$. Magnification has a magnitude of $2/3$. The image of the wire is a square of side $(2/3) \times 3 \text{ cm} = 2.0 \text{ cm}$, i.e., area 4.0 cm^2 .
- 10.5** 1.33; $(9.4 - 7.7) \text{ cm} = 1.7 \text{ cm}$

10.6 $n_{ga} = \frac{\sin 60^\circ}{\sin 35^\circ} = 1.51$

$$n_{wa} = \frac{\sin 60^\circ}{\sin 41^\circ} = 1.32$$

$$\frac{\sin 45^\circ}{\sin r} = n_{gw} = \frac{n_{ga}}{n_{wa}} = 1.144$$

which gives $\sin r = 0.6181$ i.e., $r \approx 38^\circ$

- 10.7** If r is the radius (in m) of the largest circle from which light comes out and i_c is the critical angle for water-air interface, $r = 0.8 \times \tan i_c$ and $\sin i_c = 1/1.33 \approx 0.75$

$$\begin{aligned} \text{Area} &= \frac{\pi \times (0.8)^2 \times (0.75)^2}{1 - (0.75)^2} \text{ m}^2 \\ &= 2.6 \text{ m}^2 \end{aligned}$$

10.8 $n = \frac{\sin[(A + D_m)/2]}{\sin[A/2]}$

$$= \frac{\sin 50^\circ}{\sin 30^\circ}$$

$$= 2 \times 0.766 = 1.532 \approx 1.53$$

$$\frac{\sin[(30^\circ + D'_m)/2]}{\sin 30^\circ} = \frac{1.53}{1.33}$$

which gives $D'_m \approx 10^\circ$

- 10.9** Image formed on a screen is real. The lens must be a converging lens. From the lens equation, $f = 30 \text{ cm}$. Size of the image = 10 cm.

10.10 Use the lens maker's formula to obtain $R = 22$ cm.

10.11 Here the object is virtual and the image is real. $u = +12$ cm. (object on right; virtual).

(a) $f = +20$ cm

$$\frac{1}{v} = \frac{1}{f} + \frac{1}{u} = \frac{1}{20} + \frac{1}{12}$$

$$= \frac{2}{15}$$

i.e., $v = 7.5$ cm,

(image on right; real), 7.5 cm from the lens.

(b) $f = -16$ cm

$$\frac{1}{v} = -\frac{1}{16} + \frac{1}{12} = \frac{1}{48}$$

i.e., $v = 48$ cm (image on right; real). 4.8 cm from the lens.

10.12 Image is erect, virtual and located 8.4 cm from the lens on the same side as the object. It is diminished to a size $= (8.4/14) \times 3$ cm = 1.8 cm. As the object is moved away from the lens, the virtual image moves towards the focus of the lens (but never beyond), and progressively diminishes in size.

Note that when the object is placed at the focus of the concave lens (21 cm), the image is located at 10.5 cm (not at infinity as one might wrongly think). A *virtual* object at the focus of a concave lens produces an image at infinity.

10.13 A diverging lens of focal length 60 cm

10.14 6

10.15 (a) $v_e = -25$ cm and $f_e = 6.25$ cm give $u_e = -5$ cm; $v_o = (15 - 5)$ cm = 10 cm,

$$f_o = u_o = -2.5$$
 cm; Magnifying power $= \frac{10}{2.5} \times \frac{25}{5} = 20$

(b) $u_e = -6.25$ cm, $v_o = (15 - 6.25)$ cm = 8.75 cm, $f_o = 2.0$ cm. Therefore, $u_o = -(70/27) = -2.59$ cm.

Magnifying power

$$= \frac{v_o}{|u_o|} \times (25/6.25)$$

$$= \frac{27}{8} \times 4 = 13.5$$

10.16 Angular magnification of the eye-piece for image at 25 cm

$$= \frac{25}{2.5} + 1 = 11;$$

$$|u_e| = \frac{25}{11} \text{ cm} = 2.27 \text{ cm.}$$

$$\text{Now, } \frac{1}{v_o} + \frac{1}{0.9} = (1/0.8)$$

i.e., $v_o = 7.2$ cm

Separation = $(7.2 + 2.27)$ cm = 9.47 cm

Magnifying power

$$= 11 \times \frac{7.2}{0.9} = 88$$

10.17 24; 150 cm

10.18 (a) Angular magnification

$$= \frac{15}{0.01} = 1500$$

(b) If d is the diameter of the image (in cm),

$$\frac{d}{1500} = \frac{3.48 \times 10^6}{3.8 \times 10^8}$$

i.e., $d = 13.7$ cm

10.19 Data on focal length not needed. The resolving power of the telescope is determined by

$$\theta = \frac{1.22\lambda}{D} = \frac{1.22 \times 6 \times 10^{-7}}{0.60}$$

$$= 1.22 \times 10^{-6} \text{ rad}$$

where the value of λ chosen corresponds roughly to the wavelength of yellow light. Now the transverse separation between the two sources subtends an angle equal to

$$\frac{10^{10} \text{ m}}{9.46 \times 10^{19} \text{ m}} \approx 10^{-10} \text{ rad,}$$

where we have used 1 light year = 9.46×10^{15} m. This angle is much too small compared to θ above. The two stars of the binary can not be resolved by the given telescope.

10.20 (a) 5

(b) Distant images arise due to multiple reflections. At each reflection, part of the incident intensity of light is lost due to absorption etc.

(c) For search light, a parallel beam of light is required. A divergent beam is not useful because then intensity diminishes with distance (recall inverse square dependence). Now, in case of a concave spherical mirror only paraxial rays (i.e., those close to the axis) parallel to the axis are brought to a sharp focus. Conversely, a source placed at the focus of the concave mirror produces a parallel beam only close to the axis. Rays reflected from points not close to the axis give rise to a divergent beam. This is not so in case of a parabolic mirror. A source placed at the focus of a parabolic mirror produces a parallel beam of wide cross-section. Hence, its use as a search-light mirror.

(d) A convex mirror gives a much wider field of view of the traffic at your back, than a plane mirror of the same size. However, it gives an erroneous idea of the movement of the vehicles. Because of the first advantage, it is still preferred to a plane mirror.

10.21 (a) Focal length of a mirror is about half its radius of curvature and has nothing to do with the external medium. The focal length of the convex lens will increase because the refractive index of glass with respect to water is less than refractive index of glass with respect to air.

- (b) The air layers closer to the ground are hotter than higher layers. Oblique rays coming from distant sky therefore travel from denser to rarer parts of the atmosphere and get more and more oblique. When the angle of incidence exceeds critical angle (for dense air-rarer air interface), rays get totally reflected and may enter the observer's eye. The observer, therefore, sees a reflected image of the distant part of the sky.
- (c) The apparent position of a star is slightly different from the actual position due to refraction of starlight by the atmosphere. Further, this apparent position is not stationary, since the conditions of the refracting medium are not stationary. Starlight travels through fluctuating masses of air in motion with changing conditions of temperature, temperature gradients etc. The fluctuating apparent position of the star gives rise to the twinkling effect.
- (d) Since the atmosphere bends starlight towards the normal, the apparent position of a star is slightly 'above' its actual position. Thus, even when the Sun has actually set (i.e., gone below the horizon) its apparent position remains above the horizon for some time.
- 10.22** (a) A white body reflects all the light incident on it. A black body, in contrast, absorbs all of the light energy incident on it which then gets converted into heat. This is why dark dresses are used to keep warm in winter.
- (b) If an object looks blue in white light, it means that it absorbs all the colours except those in the blue region. Light from a sodium lamp is yellow which therefore is nearly wholly absorbed by the object. The object will appear black.
- (c) The mask has a filter that absorbs the ultraviolet radiation (which is dangerous for the eyes) produced by the welding arc.
- (d) Scattering of light by the atmosphere is colour-dependent. Blue light is scattered much more strongly than red light. The blue component of sunlight is therefore proportionately more in light coming from different parts of the atmosphere. This gives the impression of the blue sky. In the evening, when the Sun is near the horizon, sunlight has to travel through much greater distances than at noon. Thus, a larger proportion of the blue component of sunlight gets scattered away. Light reaching the observer therefore has larger proportion of the remaining colour. The Sun, therefore, appears orange or red.
- 10.23** (a) The camera lens is moved towards or away from the film. For instance, if a very distant object is being focussed, the distance between the lens and the film is about the focal length of the lens. At closer object distance, the camera lens must be moved away from the film to provide the required greater image distance. In practice, because of the small focal length of the lens in an ordinary camera (about 5 cm) and the large object distances involved, only a minor movement of the camera lens serves the purpose.
- (b) f -number of a camera lens describes the ratio of focal length to the diameter of the aperture of the lens. Thus, $f/11$ means the diameter of the aperture is focal length divided by 11. Now, the capacity to collect light depends on the square of the aperture size. Thus, the given sequence of apertures (for a fixed focal length) have a light gathering capacity proportional to

$$\frac{1}{2^2}, \frac{1}{(2.8)^2}, \frac{1}{4^2}, \frac{1}{(5.6)^2}, \frac{1}{8^2}, \frac{1}{11^2}$$

which is roughly in the ratio

$$1: \frac{1}{2} : \frac{1}{4} : \frac{1}{8} : \frac{1}{16} : \frac{1}{32}$$

The corresponding exposure times required to receive the same total amount of light are in the ratio: 2 : 4 : 8 : 16 : 32.

- (c) The shutter controls the exposure time, usually to convenient steps such as (1/60)s, (1/20)s etc. For a proper photograph, the total amount of light received by the film should be within certain limits. The aperture size and exposure time together determine the total amount of light received.

10.24 The reflected rays get deflected by twice the angle of rotation of the mirror. Therefore, $d/1.5 = \tan 7^\circ$ i.e., $d = 18.4$ cm.

10.25 For the head to be seen by the eye, the top edge of the mirror should not be lower than 1.44 m from the ground. For the foot to be seen, the bottom edge of the mirror should not be higher than 0.69 m from ground. Thus, the minimum length of the mirror for a full view is $(1.44 - 0.69)$ m = 0.75 m. This minimum length is the same for any level, but the positions of the top and bottom edges of the mirror will depend on the eye-level.

10.26 Time required by light for the round trip M to N to M

$$= \frac{4.5 \times 10^4}{3 \times 10^8} = 1.5 \times 10^{-4} \text{ s}$$

During this time the mirror rotates by $(1/2) \times 27^\circ = 13.5^\circ$. Speed of rotation of the mirror

$$= \frac{13.5}{360} \times \frac{1}{10^{-4} \times 1.5} \text{ rev/s}$$

$$= 250 \text{ rev s}^{-1}$$

10.27 (a) $\frac{1}{v} + \frac{1}{u} = \frac{1}{f}$ or $\frac{1}{v} = \frac{1}{f} - \frac{1}{u}$

$f < 0$ (concave mirror);

$u < 0$ (object on left)

For $2f < u < f$ implies

$$\frac{1}{2f} > \frac{1}{u} > \frac{1}{f}$$

or $-\frac{1}{2f} < -\frac{1}{u} < -\frac{1}{f}$

or $\frac{1}{f} - \frac{1}{2f} < \frac{1}{f} - \frac{1}{u} < 0$

or $\frac{1}{2f} < \frac{1}{v} < 0$

which means $v < 0$ (image on left, real), the image lies beyond $2f$. The image is real because v is negative.

(b) $\frac{1}{v} = \frac{1}{f} - \frac{1}{u}$. Now for a convex mirror $f > 0$. Also we have $u < 0$ (object

on left). Therefore, $(1/v)$ or $v > 0$ (image on right; virtual) i.e., the image is virtual whatever be the value of u .

(c) $\frac{1}{v} = \frac{1}{f} - \frac{1}{u}$. Since, $f > 0$ (convex mirror) and $u < 0$, $(1/v) > (1/f)$ i.e., v

$< f$ (image located between the pole and the focus). And from the above, $v < |u|$ (image diminished).

(d) $\frac{1}{v} = \frac{1}{f} - \frac{1}{u}$; $f < 0$ (concave mirror), $f < u < 0$ implies $\frac{1}{f} - \frac{1}{u} > 0$,

i.e., $(1/v) > 0$ or $v > 0$ (image on right; virtual).

Also $\frac{1}{v} < \frac{1}{|u|}$ i.e., $v > |u|$ (image enlarged).

10.28 The pin appears raised by $15 [1 - (1/1.5)]$ cm = 5.0 cm. You can see from an explicit ray diagram that the answer is independent of the location of the slab (for small angles of incidence).

10.29 Use the important result of Exercise 10.28. The displacement between the virtual image due to refraction and the object depends only on the thickness and refractive index of the intervening medium. If a second medium is interposed, the image due to the first becomes an object for the second medium. The total displacement between the final image and the object is then the sum of the displacements due to each medium. Thus, total displacement

$$= 4 \left(1 - \frac{1}{1.5} \right) \text{cm} + 6 \left(1 - \frac{1}{1.4} \right) \text{cm} + 8 \left(1 - \frac{1}{1.3} \right) \text{cm} = 4.9 \text{cm}$$

10.30 You will see that rays inside the prism undergo total internal reflections, twice (once on each side of the prism) in case (a); and once (on the hypotenuse of the prism) in case (b), provided the refracted ray in the prism meets the hypotenuse.

10.31 (a) $\sin i'_c = 1.44/1.68$ which gives $i'_c = 59^\circ$. Total internal reflection takes place when $i > 59^\circ$ or when $r, < r_{\max} = 31^\circ$. Now, $(\sin i_{\max} / \sin r_{\max}) = 1.68$, which gives $i_{\max} \simeq 60^\circ$. Thus, all incident rays of angles in the range $0 < i < 60^\circ$ will suffer total internal reflections in the pipe. (If the length of the pipe is finite, which it is in practice, there will be a lower limit on i determined by the ratio of the diameter to the length of the pipe).

(b) If there is no outer coating, $i'_c = \sin^{-1}(1/1.68) = 36.5^\circ$. Now, $i = 90^\circ$ will have $r = 36.5^\circ$ and $i' = 53.5^\circ$ which is greater than i'_c . Thus, all incident rays (in the range $0 < i < 90^\circ$) will suffer internal reflections.

10.32 The slit can be viewed by the light reflected from either face to the prism. For two parallel incident rays, one on each face, the angle between the corresponding reflected rays can be shown to be twice the angle of the prism. Therefore, $A = 72^\circ$.

10.33 Lens equation

$$\frac{1}{v} - \frac{1}{u} = \frac{1}{f}$$

(New cartesian sign convention).

- (a) Convex lens:
- $f > 0$
- ;
- $u < 0$
- (object on left).

For $0 < |u| < f$

$$\frac{1}{v} = \frac{1}{f} + \frac{1}{u}$$

$$= \frac{1}{f} - \frac{1}{|u|} < 0$$

i.e., $v < 0$ (image on left; virtual).

Also for this case

$$\frac{1}{|v|} < \frac{1}{|u|} \text{ i.e., } |v| > |u|$$

(image enlarged)

- (b) Concave lens :
- $f < 0$
- ;
- $u < 0$
- (object on left)

$$\frac{1}{v} = \frac{1}{f} + \frac{1}{u}$$

$$= -\left(\frac{1}{|f|} + \frac{1}{|u|}\right) < 0$$

for all u i.e., $v < 0$ for all u (image on left; virtual). Also

$$\frac{1}{|v|} = \frac{1}{|f|} + \frac{1}{|u|}$$

$$\text{i.e., } \frac{1}{|v|} > \frac{1}{|u|}$$

$$\text{i.e., } |v| < |u|$$

(image diminished).

- 10.34** (a) The front surface of the thick glass mirror is both reflecting and refracting. The back surface is silvered and acts as a mirror. Images arise due to reflection of the incident light by the front surface then by the back surface (the bright image), followed by multiple reflections of light within the glass by the front and back surfaces. You can draw a ray diagram to see all this more clearly.
- (b) You read because the light scattered by the newspaper causes diffuse reflection. An image is seen only when reflection is regular (or specular as it is sometimes called), i.e., when parallel incident rays are reflected in parallel direction. The surface inhomogeneities of the paper are responsible for diffuse reflection.
- (c) Rays converging to a point 'behind' a plane or convex mirror are reflected to a point in front of the mirror on a screen. In other words, a plane or convex mirror can produce a real image if the object is virtual. Convince yourself by drawing an appropriate ray diagram.
- (d) Nature exhibits left-right symmetry. That is, physical laws are identical for you and your mirror image. That is, the movements of your mirror

image are perfectly physically possible movements and if an outsider views the two films he cannot say which of the two is the mirror image film. Of course, any additional left-right *asymmetric* initial information which you may have can help decide which is which. For example, if it is known that you are, say, a left-hander, then your image will appear to be right-hander. And from watching the two films you can tell which is the mirror image film. The important thing is not to confuse this possibility of asymmetric initial condition with left-right asymmetry in physical laws, which as already said, do not distinguish between you and your mirror image. (**Note:** The left-right symmetry of nature, however, does not hold good at the sub-atomic level for some types of processes.)

- 10.35** (a) When the reflected or refracted rays are divergent, the image is virtual. The divergent rays can be converged on to a screen by means of an appropriate converging lens. The convex lens of the eye does just that. The virtual image here serves as an object for the lens to produce a real image. Note, the screen here is not located at the position of the virtual image. There is no contradiction.
- (b) Taller
- (c) The apparent depth for oblique viewing decreases from its value for near-normal viewing. Convince yourself of this fact by drawing ray diagrams for different positions of the observer.
- (d) Refractive index of a diamond is about 2.42, much larger than that of ordinary glass (about 1.5). The critical angle of diamond is about 24° , much less than that of glass. A skilled diamond-cutter exploits the larger range of angles of incidence (in the diamond), 24° to 90° , to ensure that light entering the diamond is totally reflected from many faces before getting out—thus producing a sparkling effect.

- 10.36** For fixed distance s between object and screen, the lens equation does not give a real solution for u or v if f is greater than $s/4$.

Therefore, $f_{\max} = 0.75 \text{ m}$.

- 10.37** (a) $-u$ and v interchange their values at the two locations

$$-u \text{ (or } v) = \frac{90 - 20}{2} \text{ cm} = 35 \text{ cm}$$

$$v \text{ (or } -u) = 55 \text{ cm}$$

$$f = \frac{55 \times 35}{90} \\ = 21.4 \text{ cm}$$

$$(b) \frac{4.6}{h} = \frac{v}{|u|},$$

$$\frac{1.7}{h} = \frac{|u|}{v}$$

(because $|u|$, v interchange their values at the two locations).

$$\text{Thus, } h = \sqrt{4.6 \times 1.7} \text{ cm} = 2.8 \text{ cm}.$$

- 10.38** (a) (i) Let a parallel beam be the incident from the left on the convex lens first.
- $f_1 = 30 \text{ cm}$ and $u_1 = -\infty$, give $v_1 = +30 \text{ cm}$. This image becomes a virtual object for the second lens.

$f_2 = -20$ cm $u_2 = + (30 - 8)$ cm = +22 cm. Therefore,

$$\frac{1}{v_2} = -\frac{1}{20} + \frac{1}{22}$$

which gives $v_2 = -220$ cm. The parallel incident beam appears to diverge from a point 216 cm from the centre of the two-lens system.

- (ii) Let the parallel beam be incident from the left on the concave lens first: $f_1 = -20$ cm, $u_1 = -\infty$, give $v_1 = -20$ cm. This image becomes a real object for the second lens: $f_2 = +30$ cm, $u_2 = - (20 + 8)$ cm = -28

cm. Therefore,
$$\frac{1}{v_2} = \frac{1}{30} - \frac{1}{28}$$

i.e., $v_2 = -420$ cm

The parallel incident beam appears to diverge from a point 416 cm on the left of the centre of the two-lens system.

Clearly, the answer depends on which side of the lens system the parallel beam is incident. Further we do not have a simple lens equation true for all u (and v) in terms of a definite constant of the system (the constant being determined by f_1 and f_2 , and the separation between the lenses). The notion of effective focal length, therefore, does not seem to be meaningful for this system.

- (b) $u_1 = -40$ cm, $f_1 = 30$ cm, give $(1/v_1) + (1/40) = (1/30)$ i.e., $v_1 = 120$ cm. Magnitude of magnification due to the first (convex) lens

$$= 120/40 = 3$$

$$u_2 = + (120 - 8)$$
 cm = +112 cm

(object virtual);

$$f_2 = -20$$
 cm which give

$$\frac{1}{v_2} = -\frac{1}{20} + \frac{1}{112}$$

$$\text{i.e., } v_2 = -\frac{112 \times 20}{92} \text{ cm}$$

Magnitude of magnification due to the second (concave) lens = $20/92$.

Net magnitude of magnification = $3 \times (20/92) = 0.652$

Size of the image = 0.652×1.5 cm = 0.98 cm

- 10.39** If the refracted ray in the prism is incident on the second face at the critical angle i_c , the angle of refraction r at the first face is $(60^\circ - i_c)$.

$$\text{Now, } i_c = \sin^{-1} (1/1.524) \simeq 41^\circ$$

Therefore, $r = 19^\circ$

$$\sin i = (\sin 19^\circ) \times 1.524 = 0.4962$$

$$i = \sin^{-1} 0.4962 \simeq 30^\circ$$

- 10.40** The first measurement gives the focal length f of the combination of the convex lens and the plano-convex liquid lens. The second measurement gives the focal length f_1 of the convex lens. The focal length f_2 of the plano-convex lens is then given by:

$$\frac{1}{f_2} = \frac{1}{45} - \frac{1}{30} = -\frac{1}{90}$$

$$\text{i.e., } f_2 = -90 \text{ cm}$$

Using the lens maker's formula for the equiconvex lens,

$$\frac{1}{30} = (1.5 - 1) \left(\frac{1}{R} + \frac{1}{R} \right)$$

which gives $R = 30$ cm.

The same formula applied to the plano-convex lens gives

$$-\frac{1}{90} = (n - 1) \left(\frac{1}{30} + 0 \right)$$

from which the refractive index of liquid, $n = 1.33$.

10.41 Use

$$n = \frac{\sin[(A + D_m) / 2]}{\sin[A / 2]}$$

to get

Crown-glass:

$$n_b = 1.520, n_r = 1.509, n_y = 1.516$$

Flint-glass:

$$n_b = 1.666, n_r = 1.644, n_y = 1.657$$

Use

$$\omega = \frac{n_b - n_r}{n_y - 1}$$

$$\omega_1 = 0.0213 \text{ (crown-glass)}$$

$$\omega_2 = 0.0335 \text{ (flint-glass)}$$

The ratio of dispersive power of flint-glass to crown-glass is 1.57.

10.42 Two identical prisms made of the same material placed with their bases on opposite sides (of the incident white light) and faces touching (or parallel) will neither deviate nor disperse, but will merely produce a parallel displacement of the beam.

(a) To deviate without dispersion, choose, say, the first prism to be of crown-glass, and take for the second prism a flint-glass prism of suitably chosen refracting angle (smaller than that of crown-glass prism because the flint-glass prism disperses more) so that dispersion due to the first is nullified by the second.

(b) To disperse without deviation, increase the angle of flint glass prism (i.e., try flint-glass prisms of greater and greater angle) so that deviations due to the two prisms are equal and opposite. (The flint-glass prism angle will still be smaller than that of crown-glass because flint-glass has higher refractive index than that of crown-glass). Because of the adjustments involved for so many colours, these are not meant to be precise arrangements for the purpose required.

10.43 We can take average focal length of 15 cm to correspond to yellow colour. At a point slightly less than 15 cm, the centre of the spot will be violet with a red edge and other colours in between. As the screen is moved away, the centre of the spot changes through succession of blue, green, yellow, orange and finally red. When the centre of the spot is red, the edge will be violet with other colours in between.

- 10.44** (a) Using the lens maker's formula for blue, red and yellow separately, we obtain:

$$\frac{1}{f_1^b} - \frac{1}{f_1^r} = \frac{\omega_1}{f_1^y}$$

and

$$\frac{1}{f_2^b} - \frac{1}{f_2^r} = \frac{\omega_2}{f_2^y}$$

Thus,

$$\begin{aligned} \frac{1}{f^b} - \frac{1}{f^r} &= \left(\frac{1}{f_1^b} + \frac{1}{f_2^b} \right) - \left(\frac{1}{f_1^r} + \frac{1}{f_2^r} \right) \\ &= \frac{\omega_1}{f_1^y} + \frac{\omega_2}{f_2^y} \end{aligned}$$

For an 'achromatic doublet', $f^b = f^r$ which gives

$$\frac{f_1^y}{f_2^y} = -\frac{\omega_1}{\omega_2}$$

- (b) Combine the given flint-glass lens ($f_1^y = 15$ cm) with a concave lens made of crown glass of focal length $f_2^y = -10$ cm.

- 10.45** Double convex lens (crown glass)

$$\begin{aligned} \frac{1}{f_b} - \frac{1}{f_r} &= \frac{(n_b - n_r) \times 2}{15} = (1.520 - 1.509) \times \frac{2}{15} \\ &= 0.0014666 \text{ cm}^{-1} \end{aligned}$$

Flint-glass lens

$$\begin{aligned} \frac{1}{f_b} - \frac{1}{f_r} &= (1.660 - 1.644) \times \left(-\frac{1}{15} + \frac{1}{R} \right) \\ &= -0.0014666 \text{ cm}^{-1} \end{aligned}$$

This gives $R = \infty$

The other surface of the flint-glass lens is plane; it is plano-convex lens.

- 10.46** (a) Not necessarily. A material may reflect one colour strongly and transmit another colour e.g., some lubricating oils reflect green light and transmit red light.
(b) Green with a tinge of blue and yellow.

- 10.47** To see objects at infinity, the eye uses its least converging power = $(40+20)$ dioptries = 60 dioptries. This gives a rough idea of the distance between the retina and cornea-eye-lens: $(5/3)$ cm. To focus an object at the near point ($u = -25$ cm), on the retina ($v = 5/3$ cm) the focal length should be

$$\left[\frac{1}{25} + \frac{3}{5} \right]^{-1} = \frac{25}{16} \text{ cm}$$

corresponding to a converging power of 64 dioptries. The power of the eye-lens then is $(64 - 40)$ dioptries = 24 dioptries. The range of accommodation of the eye-lens is roughly 20 to 24 dioptries.

10.48 No, a person may have normal ability of accommodation of the eye-lens and yet may be myopic or hypermetropic. Myopia arises when the eye-ball from front to back gets too shortened. In practice, in addition, the eye-lens may also lose some of its ability of accommodation. When the eyeball has the normal length but the eye-lens loses partially its ability of accommodation (as happens with increasing age for any normal eye) the 'defect' is called presbyopia and is corrected in the same manner as hypermetropia.

- 10.49** (a) Concave lens of focal length = -80 cm i.e., of power = -1.25 dioptres.
 (b) No. The concave lens, in fact, reduces the size of the object (image distance is less than object distance), but the angle subtended by the distant object at the eye is the same as the angle subtended by the image (on the far point) at the eye. The eye is able to see distant objects not because the corrective lens magnifies the object, but because it brings the object (i.e., it produces virtual image of the object) at the far point of the eye which then can be focussed by the eye-lens on the retina.
 (c) The myopic person may have a normal near point i.e., about 25 cm (or even less). In order to read a book with his spectacles (for distant vision), he must keep the book at a greater distance than 25 cm so that the image of the book by the concave lens is produced not closer than 25 cm. The angular size of the book (or its image) at the greater distance is evidently less than the angular size when the book is placed at 25 cm and no spectacles are used. Hence, the person prefers to remove his spectacles while reading.

- 10.50** (a) $u = -25$ cm, $v = -75$ cm

$$\frac{1}{f} = \frac{1}{25} - \frac{1}{75}$$

i.e., $f = 37.5$ cm.

The corrective lens has a converging power of $+2.67$ dioptres.

- (b) The corrective lens produces a virtual image (at 75 cm) of an object at 25 cm. The angular size of this image is the same as that of the object. In this sense the lens does not magnify the object but merely brings the object to the near point of the eye which then gets focussed by the eye-lens on the retina. However, the angular size is greater than that of the same object at the near point (75 cm) viewed without the spectacles.
 (c) A hypermetropic eye may have normal far point i.e., it may have enough converging power to focus parallel rays from infinity on the retina of the shortened eyeball. Wearing spectacles of converging lenses (used for near vision) will amount to more converging power than needed for parallel rays. The result will be that distant objects may get focussed in front of the retina (like a myopic eye) and will appear blurred.

- 10.51** (a) $f = -(1/0.8) \text{ m} = -125$ cm; $v = -80$ cm.

Using the lens equation, $u = -222$ cm. He can see objects up to 2.22 m clearly.

- (b) $f = +(1/1) \text{ m} = 100$ cm; $v = -75$ cm. Using the lens equation, $u = -42.9$ cm. His nearest distance of distinct vision is about 43 cm.

- 10.52** The far point of the person is 100 cm, while his near point may have been normal (about 25 cm). Objects at infinity produce virtual image at 100 cm (using spectacles). To view closer objects i.e., those which are (or whose images using the spectacles are) between 100 cm and 25 cm, the person uses the ability of accommodation of his eye-lens. This ability usually gets

partially lost in old age (presbyopia). The near point of the person recedes to 50 cm. To view objects at 25 cm clearly, the person needs converging lens of power +2 dioptries.

- 10.53** The defect (called astigmatism) arises because the curvature of the cornea plus eye-lens refracting system is not the same in different planes. [The eye-lens is usually spherical i.e., has the same curvature on different planes but the cornea is not spherical in case of an astigmatic eye]. In the present case, the curvature in the vertical plane is enough, so sharp images of vertical wires can be formed on the retina. But the curvature is insufficient in the horizontal plane, so horizontal wires appear blurred. The defect can be corrected by using a cylindrical lens with its axis along the vertical. Clearly, parallel rays in the vertical plane will suffer no extra refraction, but those in the horizontal plane can get the required extra convergence due to refraction by the curved surface of the cylindrical lens if the curvature of the cylindrical surface is chosen appropriately.

10.54 (a) Closest distance = $4\frac{1}{6}$ cm ≈ 4.2 cm $\left(\text{use: } -\frac{1}{25} - \frac{1}{u} = \frac{1}{5} \right)$

Farthest distance = 5 cm $\left(\text{use: } -\frac{1}{\infty} - \frac{1}{u} = \frac{1}{5} \right)$

- (b) Maximum angular magnification = $[25/(25/6)] = 6$. Minimum angular magnification = $(25/5) = 5$

10.55 (a) $\frac{1}{v} + \frac{1}{9} = \frac{1}{10}$

i.e., $v = -90$ cm,

Magnitude of magnification = $90/9 = 10$.

Each square in the virtual image has an area $10 \times 10 \times 1 \text{ mm}^2 = 100 \text{ mm}^2 = 1 \text{ cm}^2$

- (b) Magnifying power = $25/9 = 2.8$

(c) No, magnification of image by a lens and angular magnification (or magnifying power) of an optical instrument are two separate things. The latter is the ratio of the angular size of the object (which is equal to the angular size of the image even if the image is magnified) to the angular size of the object if placed at the near point (25 cm). Thus, magnification magnitude is $|v/u|$ and magnifying power is $(25/|u|)$. Only when the image is located at the near point $|v| = 25$ cm, are the two quantities equal.

- 10.56** (a) Maximum magnifying power is obtained when the image is at near the point (25 cm)

$$-\frac{1}{25} - \frac{1}{u} = \frac{1}{10}$$

i.e., $u = -\frac{50}{7} = -7.14$ cm.

- (b) Magnitude of magnification = $(25/|u|) = 3.5$.

- (c) Magnifying power = 3.5

Yes, the magnifying power (when the image is produced at 25 cm) is equal to the magnitude of magnification. See answer to Exercise 10.55.

$$10.57 \text{ Magnification} = \sqrt{6.25/1} = 2.5$$

$$v = +2.5u$$

$$+\frac{1}{2.5u} - \frac{1}{u} = \frac{1}{10}$$

$$\text{i.e., } u = -6 \text{ cm}$$

$$|v| = 15 \text{ cm}$$

The virtual image is closer than the normal near point (25 cm) and cannot be seen by the eye distinctly.

- 10.58** (a) Even though the absolute image size is bigger than the object size, the angular size of the image is equal to the angular size of the object. The magnifier helps in the following way: without it object would be placed no closer than 25 cm; with it the object can be placed much closer. The closer object has larger angular size than the same object at 25 cm. It is in this sense that angular magnification is achieved.
- (b) Yes, it decreases a little because the angle subtended at the eye is then slightly less than the angle subtended at the lens. The effect is negligible if the image is at a very large distance away. [Note: When the eye is separated from the lens, the angle subtended at the eye by the first object and its image is not equal].
- (c) First, grinding lens of very small focal length is not easy. More important, if you decrease focal length, aberrations (both spherical and chromatic) become more pronounced. So, in practice, you cannot get a magnifying power of more than 3 or so with a simple convex lens. However, using an aberration corrected lens system, one can increase this limit by a factor of 10 or so.
- (d) Angular magnification of eye-piece is $(25/f_e) + 1$ (f_e in cm) which increases if f_e is smaller. Further magnification of the objective is given

$$\text{by } \frac{v_o}{|u_o|} = \frac{1}{(|u_o|/f_o) - 1}$$

which is large when $|u_o|$ is slightly greater than f_o . The microscope is used for viewing very close object. So $|u_o|$ is small, and so is f_o .

- (e) The image of the objective of the eye-piece is known as 'eye -ring'. All the rays from the object refracted by objective go through the eye-ring. Therefore, it is an ideal position for our eyes for viewing. If we place our eyes too close to the eye-piece, we shall not collect much of the light and also reduce our field of view. If we position our eyes on the eye-ring and the area of the pupil of our eye is greater or equal to the area of the eye-ring, our eyes will collect all the light refracted by the objective. The precise location of the eye ring naturally depends on the separation between objective and eye-piece. When you view through a microscope by placing your eyes on one end, the ideal distance between the eyes and eye-piece is usually built in the design of the instrument.

- 10.59** Assume microscope in normal use i.e., image at 25 cm. Angular magnification of the eye-piece

$$= \frac{25}{5} + 1 = 6$$

Magnification of the objective

$$= \frac{30}{6} = 5$$

$$\frac{1}{5u_0} - \frac{1}{u_0} = \frac{1}{1.25}$$

which gives $u_0 = -1.5$ cm; $v_0 = 7.5$ cm. $|u_e| = (25/6)$ cm = 4.17 cm. The separation between the objective and the eye-piece should be $(7.5 + 4.17)$ cm = 11.67 cm. Further the object should be placed 1.5 cm from the objective to obtain the desired magnification.

10.60 (a) $m = (f_0/f_e) = (140/5) = 28$

(b) $m = \frac{f_0}{f_e} \left[1 + \frac{f_0}{25} \right]$
 $= 28 \times 1.2 = 33.6$

10.61 (a) $f_0 + f_e = 145$ cm

(b) Angle subtended by the tower = $(100/3000) = (1/30)$ rad.
 Angle subtended by the image produced by the objective

$$= \frac{h}{f_0} = \frac{h}{140}$$

Equating the two,

$$h = \frac{14}{3} \text{ cm} = 4.7 \text{ cm.}$$

(c) Magnification (magnitude) of the eye-piece

$$= \frac{25}{5} + 1 = 6$$

Height of the final image (magnitude)

$$= \frac{14}{3} \times 6 \text{ cm} = 28 \text{ cm.}$$

10.62 $-\frac{1}{u_e} + \frac{1}{40} = \frac{1}{10}$

i.e., $|u_e| = (40/3)$ cm; magnification of the eye-piece = 3. Therefore, diameter of the image formed by the objective = $(6/3)$ cm = 2.0 cm.
 If D is the diameter of the Sun (in m),

$$\frac{D}{1.5 \times 10^{11}} = \frac{2}{100}$$

His rough estimate of the size of the Sun: $D = 1.5 \times 10^9$ m (correct value = 1.39×10^9 m).

10.63 (a) No chromatic aberration due to the objective because only reflection is involved; spherical aberration reduced by using a mirror of the shape of paraboloid; brighter image than in a refracting telescope of equivalent size because in the latter intensity of light is partially lost due to reflection and absorption by the objective lens glass; mirror entails grinding and polishing of only one side; high resolution (as well as brightness of a point object) achieved by using a mirror of large aperture which is easier

to support (its back side being available) than a lens of the same aperture.

- (b) Focal length of the mirror = - 40 cm;
Magnifying power = $40/1.6 = 25$

10.64 (a) See the answer to Exercise 10.58(e).

- (b) First, let us determine the location of the eye-ring: Distance of the objective from the eye-piece, $u = -(f_o + f_e)$
Image (i.e., eye-ring) distance = v .

$$\text{Use } \frac{1}{v} + \frac{1}{f_o + f_e} = \frac{1}{f_e}$$

$$\text{to get } v = \frac{(f_o + f_e)f_e}{f_o}$$

Linear magnification

$$\frac{v}{|u|} = \frac{f_e}{f_o}$$

$$\text{But linear magnification} = \frac{\text{Diameter of the eye-ring}}{\text{Diameter of the objective}}$$

Using angular magnification = (f_o/f_e) , the desired result is proved.

- (c) Use the result obtained in (b). Diameter of the objective = 300×0.3 cm
= 90 cm

10.65 In a terrestrial telescope, the inverted image formed by the objective is made erect by positioning it at the $2f$ point of an erecting lens of focal length f . Thus the separation between the objective and eye-piece is $f_o + f_e + 4f = (180 + 5 + 14)$ cm = 199 cm. Magnifying power remains unaltered: $(f_o/f_e) = (180/5) = 36$. The telescope can, of course, be used to view astronomical objects though obviously there is no need to make the 'inverted' image of star 'upright' (This is, however, necessary for viewing a terrestrial object). The final image obtained is somewhat less brighter than in an equivalent astronomical telescope because of the extra loss of some light due to reflection and absorption by the erecting lens.

10.66 (a) Draw a ray diagram for this telescope, remembering that the image formed by the objective is a virtual object for the eye-piece interposed between the objective and its focus (which is also the focus of the eye-piece in normal adjustment). The derivation is identical to that for an ordinary astronomical telescope.

- (b) Separation = $(150 - 7.5)$ cm = 142.5 cm

- (c) Limited field of view because the eye cannot be positioned on the location of the eye-ring between the two lenses.

10.67 Main advantage: compactness. The effective length of the telescope is three times the distance between the objective and the eye-piece.

10.68 (a) Light-gathering capacity of A is 9 times that of B.

- (b) Take a star at distance l that is barely visible in a telescope using the objective of diameter d . The intensity of light from a star (of the same absolute brightness) at a distance $2l$ is reduced by a factor of 4. To receive the same total amount of light, the diameter of the objective should be $2d$, using the result similar to (a). Thus if the distance of the

star is doubled, the diameter of the objective should be doubled for the star to be again barely visible. Thus the distance a telescope can penetrate through the sky (range) is proportional to the diameter of its objective.

- (c) Background is not like a point source but is like an object of finite size. If m is the magnifying power, the area of the image seen in the telescope is m^2 the area seen by the unaided eye. But m is also the ratio of the diameter of the objective to the diameter of the eye-ring. Assuming our eye pupil fills up the eye-ring, it is clear that the light received by the objective is m^2 the light received by our unaided eye. Thus, for an object of finite extension, the telescope increases the area of the image seen and the total light received in the same proportion. The net effect is that the amount of light per unit area i.e., brightness is not increased. It is, in fact, slightly reduced due to absorption etc., of some light by the various optical elements of the telescope. The argument does not go through for a point source, so that the brightness of a point source does increase using a telescope. A star, therefore, appears brighter against its background when viewed through a telescope.

10.69 Linear magnification

$$= \frac{1000}{24} = \frac{125}{3}$$

$$-(u + \frac{125}{3}u) = 12$$

$$\text{i.e., } u = \frac{-9}{32} \text{ m,}$$

$$v = \frac{375}{32} \text{ m}$$

$$\text{Using } \frac{1}{v} - \frac{1}{u} = \frac{1}{f}$$

$$f = 27.5 \text{ cm.}$$

The projection lens should be 28.1 cm from the slide and have focal length = 27.5 cm

- 10.70** (a) A strong illumination of the slide is necessary in a projector, otherwise the highly enlarged image on the screen will be very dim. If the slide is illuminated directly by a lamp, light diverging out of the outer portion of the slide will not be collected by the projecting lens (unless it is impractically big). We will then see only a small central part of the slide picture on the screen. A condensing lens helps converge light on the slide, thus enabling the whole of the slide to be imaged on the screen. Another important function of the condensing lens is explained in (c).
- (b) The slide is placed very close to the condensing lens in order to utilize nearly all of the light coming out of the lens for illuminating the slide. The diagonal of the slide in Exercise 10.69 is about 4.33 cm. The diameter of the condensing lens should be obviously at least about 4.4 cm in order to illuminate the whole of the slide.
- (c) An important thing in a projector is to avoid imaging the source itself on the screen. The focal length of the condensing lens should be so chosen that the image of the source is formed on the projection lens itself. Further, other parameters such as the size of the source, the diameter

of the projection lens and its placement should be such that the area of the image of the source should be about equal to the surface area of the projection lens so that light illuminating the slide is not wasted on the one hand, and on the other hand the full area of the projection lens is used for imaging the slide.

- (d) Remember if two rays originating from different parts of an object converge to a common point (after passing through some lens etc), that common point is not the image of any thing. An image is where different rays from the same point of an object converge to. The image of the source is formed on the projection lens not on the slide.
- (e) Because the image of the slide on the screen is real and inverted the 'lower' portion of the image corresponds to 'upper' portion of the slide. But each point of the slide is illuminated by all points of the source. Thus, every portion of the screen image gets illumination from all parts of the source.

- 10.71** (a) Image of a point object on a film (or a screen) is never strictly a point but is a small circle (called the circle of confusion) due to the lens aberration which can be reduced but never completely removed. Since aberrations reduce with decrease in aperture, the circle of confusion is smaller, the smaller the aperture of the lens. An object that is not properly focussed on the screen will naturally produce a large circle of confusion than the one that is properly focussed. Now, there is a limit up to which our eye can perceive the size of the small circle. Below a certain angular size ($\sim 10^{-3}$ rad), our eye perceives a circle as a sharp point. Thus, there is a *range of object distance* from the camera for which the images on the film have small enough circles of confusion for our eye to perceive them as sharp points. This range is called 'depth of field'. From what has been said above the depth of the field will clearly increase if the aperture is reduced. For photographing a scenery, we will naturally like to have a greater depth of field than for an identity photograph.
- (b) Field of view or the angle of view is determined by the size of the screen (i.e., the film) and the focal length of the camera lens [it is roughly (film size/ focal length) in radians]. To increase the field of view, one uses a lens of considerably smaller focal length than a normal camera lens (usually half of two-thirds of the normal focal length of 5 cm). This lens is known as '*wide-angle lens*'.
 - (c) A telephoto lens is the opposite of wide-angle lens. It has a considerably larger focal length than a normal camera lens and, therefore, has a much reduced field of view. Thus it photographs part of the object, but since image distance now is more, that part looks more magnified. One feels that the object has been shot from a close distance.
 - (d) A telescope views large objects at large distances; a microscope views small objects at small distances. Both need a small field of view. A camera views objects of ordinary sizes at fairly close distances. Here the field of view is required much more (compare 45° for a camera with about 1° for a microscope objective and something similar for a telescope, a moon subtends about 0.5° at the Earth). Thus rays entering a camera lens are far from being paraxial and aberrations will be large and images will be blurred if the apertures are not very small. For a telescope, on the other hand, the important thing is its ability to resolve distant objects (i.e., see them as distinct). We have seen that the resolving power increases with increase in aperture. Therefore, telescopes have as large an aperture as feasible.

- 10.72** (a) In a camera, image distance is nearly equal to the focal length of the lens (because $|u|/f \gg 1$). Therefore, linear magnification is proportional to f . Therefore, amount of light per unit area of the image is proportional to $1/f^2$. Amount of light is also proportional to exposure time t and to the square of aperture size: α^2 . Thus, the brightness of the image is proportional to $\alpha^2 t / f^2$. Therefore, for a given brightness, $t \propto f^2 / \alpha^2$ (Brightness is luminous flux per unit area).
- (b) The aperture size $f/5.6$ is about $\sqrt{2}$ times the aperture size $f/8$, i.e., the aperture $f/5.6$ gathers twice the amount of light per second gathered by the aperture $f/8$. To gather the same total amount of light, exposure time of $f/5.6$ should be half that of $f/8$ i.e., exposure time $(1/120)$ s.
- 10.73** The virtual image of the object (located at the focus of the objective) produced by L_1 should be located at the focus of L_2 . Since $L_1, L_2 = (2f/3)$, image distance from L_1 is: $v = -(f/3)$.
- Using $-\frac{1}{u} + \frac{1}{v} = \frac{1}{f}$
- $u = -(f/4)$. L_1 should be placed at a distance of $f/4$ from the focus of objective.

Chapter 11

- 11.1** (a) Spherical
(b) Plane
(c) Plane (a small area on the surface of a large sphere is nearly planar).
- 11.2** 5000 \AA , $6 \times 10^{14} \text{ Hz}$; 45°
- 11.3** (a) $2.0 \times 10^8 \text{ m s}^{-1}$ (use $v = \frac{c}{n}$)
(b) No. The refractive index, and hence the speed of light in a medium, depends on wavelength. [When no particular wavelength or colour of light is specified, we may take the given refractive index to refer to yellow colour]. Now we know violet colour deviates more than red in a glass prism, i.e. $n_v > n_r$. Therefore, the violet component of white light travels slower than the red component.
- 11.4** 400 nm , $5 \times 10^{14} \text{ Hz}$, $2 \times 10^8 \text{ m s}^{-1}$
- 11.5** Incoherent
- 11.6** $\lambda = \frac{1.2 \times 10^{-2} \times 0.28 \times 10^{-3}}{4 \times 1.4} = 600 \text{ nm}$
- 11.7** $\tan^{-1}(1.5) \simeq 56.3^\circ$
- 11.8** The factors by which the electric field gets multiplied (after the first polaroid) are $\cos 45^\circ$ and $\cos 45^\circ$ for the second and third. Multiplying these and squaring, the fractional intensity transmitted after the first polaroid is $\frac{1}{4}$.
- 11.9** 40 m
- 11.10** 12.5 cm

11.11 Use the formula

$$\lambda' - \lambda = \frac{v}{c} \lambda$$

$$\text{i.e., } v = \frac{c}{\lambda} (\lambda' - \lambda)$$

$$= \frac{3 \times 10^8 \times 15}{6563}$$

$$= 6.86 \times 10^5 \text{ m s}^{-1}$$

11.12 (a) Reflected light: (wavelength, frequency, speed same as incident light)

$$\lambda = 589 \text{ nm}, v = 5.09 \times 10^{14} \text{ Hz}, c = 3.00 \times 10^8 \text{ m s}^{-1}$$

(b) Refracted light: (frequency same as the incident frequency) $v = 5.09 \times 10^{14} \text{ Hz}$
 $v = (c/n) = 2.26 \times 10^8 \text{ m s}^{-1}$, $\lambda = (v/v) = 444 \text{ nm}$

11.13 In Newton's corpuscular (particle) picture of refraction, particles of light incident from a rarer to a denser medium experience a force of attraction normal to the surface. This results in an increase in the normal component of the velocity but the component along the surface is unchanged. This means

$$c \sin i = v \sin r \quad \text{or} \quad \frac{v}{c} = \frac{\sin i}{\sin r} = n. \text{ Since } n > 1, v > c.$$

The prediction is opposite to the experimental results ($v < c$). The wave picture of light is consistent with the experiment.

11.14 With the point object at the centre, draw a circle touching the mirror. This is a plane section of the spherical wavefront from the object that has just reached the mirror. Next draw the locations of this same wavefront after a time t in the presence of the mirror, and in the absence of the mirror. You will get two arcs symmetrically located on either side of the mirror. Using simple geometry, the centre of the reflected wavefront (the image of the object) is seen to be at the same distance from the mirror as the object.**11.15** (a) The speed of light in vacuum is a universal constant independent of all the factors listed and anything else. In particular, note the surprising fact that it is independent of the relative motion between the source and the observer. This fact is a basic axiom of Einstein's special theory of relativity.

(b) Dependence of the speed of light in a medium:

(i) does not depend on the nature of the source (wave speed is determined by the properties of the medium of propagation. This is also true for other waves, e.g., sound waves, water waves etc.).

(ii) independent of the direction of propagation for *isotropic media*.

(iii) independent of the motion of the source relative to the medium but depends on the motion of the observer relative to the medium.

(iv) depends on wavelength.

(v) independent of intensity. [For high intensity beams, however, the situation is more complicated and need not concern us here.]

11.16 Sound waves require a medium for propagation. Thus even though the situations (i) and (ii) may correspond to the same relative motion (between the source and the observer), they are not identical physically since the motion of the observer relative to the medium is different in the two situations. Therefore, we cannot expect Doppler formulas for sound to be identical for

(i) and (ii). For light waves in vacuum, there is clearly nothing to distinguish between (i) and (ii). Here only the relative motion between the source and the observer counts and the relativistic Doppler formula is the same for (i) and (ii). For light propagation in a medium, once again like for sound waves, the two situations are not identical and we should expect the Doppler formulas for this case to be different for the two situations (i) and (ii).

- 11.17** (a) Reflection and refraction (scattering in general) arise through interaction of incident light with the atomic constituents of matter. Atoms may be viewed as oscillators, which take up the frequency of the external agency (light) causing the forced oscillations. The frequency of light emitted by a charged oscillator equals its frequency of oscillation. Thus the frequency of scattered light equals the frequency of incident light. The fact can also be explained more mathematically without using the atomic picture. At any interface between the two media, the electric (and magnetic) fields must satisfy certain boundary conditions. Frequency determines the time-dependence of fields. If incident, reflected and refracted frequencies were not equal, the same boundary conditions would not be satisfied for all times.
- (b) No. Energy carried by a wave depends on the amplitude of the wave, not on the speed of wave propagation.
- (c) A pulse can be viewed as being made of harmonic waves with a large range of wavelengths. Since the speed of propagation in a medium depends on wavelength, different wavelength components of the pulse travel with different speeds. The pulse will not retain its shape as it travels through the medium.
- (d) For a given frequency, intensity of light in the photon picture is determined by the number of photons per unit area.
- (e) The speed of light in water is not independent of the relative motion between the observer and the medium. We might expect the answer to be $(c/n) + v$. The correct answer according to special relativity (and experiments) is
- $$(c/n) + v [1 - (1/n^2)] \text{ for } v \ll c.$$

- 11.18** (a) Angular separation of the fringes remains constant ($= \lambda/d$). The actual separation of the fringes increases in proportion to the distance of the screen from the plane of the two slits.
- (b) The separation of the fringes (and also angular separation) decreases. See, however, the condition mentioned in (d) below.
- (c) The separation of the fringes (and also angular separation) decreases. See, however, the condition mentioned in (d) below.
- (d) Let s be the size of the source and S its distance from the plane of the two slits. For interference fringes to be seen, the condition $s/S < \lambda/d$ should be satisfied; otherwise, interference patterns produced by different parts of the source overlap and no fringes are seen. Thus, as S decreases (i.e., the source slit is brought closer), the interference pattern gets less and less sharp, and when the source is brought too close for this condition to be valid, the fringes disappear. Till this happens, the fringe separation remains fixed.
- (e) Same as in (d). As the source slit width increases, fringe pattern gets less and less sharp. When the source slit is so wide that the condition $s/S \leq \lambda/d$ is not satisfied, the interference pattern disappears.
- (f) The angular size of the central diffraction band due to each slit is about λ/S' where S' is the width of each of the two slits. S' should be sufficiently small so that these bands are wide enough to overlap and thus produce interference. This means $\lambda/S' \gg \lambda/d$, i.e., the width of each slit should be considerably smaller than the separation between the slits. When

the slits are so wide that this condition is not satisfied, fringes are not seen. However, increase in the width of the slits does improve the brightness of the fringes. Thus, in practice, the two slits should be wide enough to allow sufficient light to pass through but narrow enough to cause enough diffraction from each slit to enable wavefront from the two slits to overlap and interfere.

- (g) The interference patterns due to different component colours of white light overlap (incoherently). The central bright fringes for different colours are at the same position. Therefore, the central fringe is white. Since blue colour has the lower λ , the fringe closest on either side of the central white fringe is blue; the farthest is red. After a few fringes, no clear fringe pattern is seen.

- 11.19** (a) Light waves coming directly from the source S and the reflected waves (which appear to come from the image S') interfere to produce a fringe pattern.
 (b) To ensure that the separation between the two coherent sources S and S' is small, as required in a Young's double slit experiment.
 (c) Reflection by mirror causes a phase change of 180° which is equivalent to a change in path length by half a wavelength.

- 11.20** (a) Let the separation between the plates at a distance x from the joining line be y

$$y = \frac{x \times S}{l}$$

Now the condition for destructive interference [see (b)] is:

$$2y = n\lambda, \quad \text{i.e.,} \quad x = \frac{l}{2S} \times n\lambda$$

Therefore, the separation between the fringes is:

$$\Delta x = \frac{l\lambda}{2S} \Delta n = \frac{l\lambda}{2S} \quad (\text{for } \Delta n = 1).$$

- (b) Reflection from the upper surface of the wedge causes no phase change (denser to rarer medium); reflection from the lower surface of the wedge causes a phase change of 180° (rarer to denser medium). Hence the fringe along the line of contact is dark.
 (c) Glass is denser than water. So the line of contact is still a dark fringe. But wavelength in water is less than that in air by a factor of 1.33. So the fringe separation is reduced by this factor.
 (d) Choose the upper plate material, medium filling the wedge, and the lower plate material in increasing (or decreasing) order of refractive index.

- 11.21** (a) Straight lines parallel to the slits.
 (b) Straight lines parallel to the line of contact of the plates forming the air wedge.
 (c) Straight lines parallel to the slits.
 (d) Here the two coherent sources are the images of the lamp (approximately a point source) in the front and back surfaces of the sheet. They produce circular fringes.
 (e) Concentric circular fringes (Newton's rings) with the centre at the point of contact of the lens and the plate.

- 11.22** (a) The angular separation between the central bright band and the first dark band is λ/d . The angular separation between the two dark bands on either side of the central bright band is therefore $2\lambda/d$. Therefore, the actual separation between the two dark bands is $(2\lambda/d) \times D$ (D : distance of the screen). For the given data, this equals 4.68 mm.

- (b) A circular hole produces circular diffraction fringes. The angular separation between the central bright band and the first dark band in this case is $1.22 \lambda/d$ (Take this result without proof). The answer, therefore, modifies to $1.22 \times 4.68 = 5.71$ mm.

- 11.23** (a) The size reduces by half according to the relation: size $\sim \lambda/d$. Intensity increases four fold.
 (b) The intensity of interference fringes in a double slit arrangement is modulated by the diffraction pattern of each slit.
 (c) Waves diffracted from the edge of the circular obstacle interfere constructively at the centre of the shadow producing a bright spot.
 (d) For diffraction or bending of waves by obstacles/apertures by a large angle, the size of the latter should be comparable to wavelength. If the size of the obstacle/aperture is much too large compared to wavelength, diffraction is by a small angle. Here the size is of the order of a few metres. The wavelength of light is about 5×10^{-7} m, while sound waves of, say, 1 kHz frequency have wavelength of about 0.3 m. Thus, sound waves can bend around the partition while light waves cannot.
 (e) Justification based on what is explained in (d). Typical sizes of apertures involved in ordinary optical instruments are much larger than the wavelength of light.

- 11.24** (a) Interference of the direct signal received by the antenna with the (weak) signal reflected by the passing aircraft.
 (b) Light waves reflected from the upper and lower surfaces of a thin film interfere. Since the condition for constructive or destructive interference (bright or dark fringe) is wavelength dependent, coloured fringes are observed.
 (c) Light wave reflected from the upper surface of the air film (denser to rarer medium) suffers a phase change of 180° . Therefore, the central fringe in the reflected light is dark. Transmitted light suffers no phase change at either surface. Hence, the central fringe in the transmitted light is bright.
 (d) If for a particular colour, interference is constructive in the reflected light, it is destructive in the transmitted light and vice versa. This is due to 180° change of phase in one of the reflected waves as mentioned in (c). Thus, if a particular colour produces a bright fringe in the reflected light, this colour will be reduced in the transmitted light from the same point. The coloured patterns in the reflected and the transmitted light, are therefore, complementary.
 (e) Superposition principle follows from the linear character of the (differential) equation governing wave motion. If y_1 and y_2 are solutions of the wave equation, so is any linear combination of y_1 and y_2 . When the amplitudes are large (e.g., high intensity laser beams) and non-linear effects are important, the situation is far more complicated and need not concern us here.

- 11.25** The electric field components in the two sets of axes are related by

$$E_x = E'_x \cos \theta - E'_y \sin \theta$$

$$E_y = E'_x \sin \theta + E'_y \cos \theta$$

Substituting for E'_x and $E'_y \propto \cos \omega t$ and $\sin \omega t$,

$$E_x = E_0 \cos (\omega t + \theta),$$

$$E_y = E_0 \sin (\omega t + \theta)$$

These describe circularly polarised light with a phase change of θ . Changing the sign of E_y is equivalent to reflecting the electric vector in the x -axis. This changes the sense of circular.

- 11.26** This is just a restatement of the previous problem. Since the E_y components have opposite signs for opposite circular polarisations, they cancel, leaving linear polarisation along x . If we want linear polarisation along x' , we should use $E'_x \propto \cos \omega t$, $E'_y \propto \pm \sin \omega t$ to build the two circular waves. Coming back to x and y components, one circularly polarised wave is shifted in phase by $+\theta$ and other by $-\theta$. The rotation of linearly polarised waves by sugar solution can be thought of as a difference in refractive index between the two opposite circular waves, producing a phase difference between them.
- 11.27** The angle between successive polaroids is $\pi/2N$. The fractional intensity is $[\cos(\pi/2N)]^{2N}$. Trying out larger values of N will quickly convince you that this approaches 1 for large N ! A mathematical expression for large N requires (i) an approximate expression for $\cos \theta$ for small θ (ii) the definition of the exponential function. If you know these, you can write, for large N ,

$$\left[1 - \frac{1}{2} \left(\frac{\pi}{2N}\right)^2\right]^{2N} \approx \exp\left(-\frac{\pi^2}{4N}\right)$$

- 11.28** (a) Changing the sign of E_y relative to E_x reflects the polarisation in the x -axis, we get linear polarisation along $-\theta$.
 (b) The sense of circular polarisation is reversed.
- 11.29** The visibility of the fringes is poorest when the path difference p is an integral multiple of λ_1 and a half integral multiple of λ_2 (for example). As p is increased, this happens first when

$$\frac{p}{\lambda_1} - \frac{p}{\lambda_2} = \frac{1}{2}$$

$$p = \frac{1}{2} \left(\frac{\lambda_1 \lambda_2}{\lambda_2 - \lambda_1} \right) = 0.29 \text{ mm.}$$

- 11.30** Divide the single slit into n smaller slits of width $a' = a/n$. The angle $\theta = n\lambda/a = \lambda/a'$. Each of the smaller slits sends zero intensity in the direction θ . The combination gives zero intensity as well.
- 11.31** The diffraction angle λ/a causes a spreading of $(L\lambda/a)$ in the size of the spot. This becomes large when a is small. Adding the two kinds of spreading (for simplicity, this is not strictly true!), we get a spot size $a + (L\lambda/a)$. To find the minimum value of this, write it as $\sqrt{[a - (L\lambda/a)]^2 + 4L\lambda}$. The minimum is when $a = (L\lambda/a)$, i.e. the geometric and diffraction broadening are equal. The minimum value is $2\sqrt{L\lambda}$.
- 11.32** Let the beams make angles $\pm (1/2)\theta$ with the normal to the screen. Moving parallel to the screen by a distance x brings one closer to the source of one beam by $x\theta/2$ and further from the source of the other by $x\theta/2$. (Draw a figure!) This causes a change in $x\theta$ in the path difference. Equating this to λ gives $x = \lambda/\theta$ for the spacing of two bright fringes. Using the film as a grating in first order gives a change in direction of $(\lambda/x) = [\lambda/(\lambda/\theta)] = \theta$. This was the angle between the original two beams. Illuminating the film with one of the two beams used to make it 'brings out' the other, inclined at θ . This problem illustrates the principle of holography. Light from a laser illuminates both, some object and a photographic film. The light scattered from the object interferes with the direct light and gives a permanent record

on the film. The developed film is used as a grating using the laser beam alone. The wavefront emerging from the grating (called a hologram) includes a copy of the wave scattered by the original object. To the eye, it presents the full three dimensional appearance of that object. The person who first suggested this technique, D. Gabor, was awarded the Nobel Prize for Physics.

Chapter 12

- 12.1** (a) 7.24×10^{18} Hz (b) 0.041 nm
- 12.2** (a) $0.34 \text{ eV} = 0.54 \times 10^{-19} \text{ J}$ (b) 0.34 V (c) 344 km/s
- 12.3** $1.5 \text{ eV} = 2.4 \times 10^{-19} \text{ J}$
- 12.4** (a) $3.14 \times 10^{-19} \text{ J}$, $1.05 \times 10^{-27} \text{ kg m/s}$ (b) 3×10^{16} photons/s (c) 0.63 m/s
- 12.5** 4×10^{21} photons/ $\text{m}^2 \text{ s}$
- 12.6** $6.59 \times 10^{-34} \text{ J s}$
- 12.7** (a) $3.38 \times 10^{-19} \text{ J} = 2.11 \text{ eV}$ (b) 3.0×10^{20} photons/s
- 12.8** (a) $5.10 \times 10^{-19} \text{ J} = 3.18 \text{ eV}$ (b) $2.62 \times 10^{-19} \text{ J} = 1.63 \text{ eV}$
- 12.9** Caesium, Potassium, Sodium
- 12.10** 2.0 V
- 12.11** No, because $v < v_0$
- 12.12** $4.73 \times 10^{14} \text{ Hz}$
- 12.13** $2.16 \text{ eV} = 3.46 \times 10^{-19} \text{ J}$
- 12.14** (a) $4.04 \times 10^{-24} \text{ kg m s}^{-1}$ (b) 0.164 nm
- 12.15** (a) $5.92 \times 10^{-24} \text{ kg m s}^{-1}$ (b) $6.50 \times 10^6 \text{ m s}^{-1}$ (c) 0.112 nm
- 12.16** (a) $6.95 \times 10^{-25} \text{ J} = 4.34 \text{ } \mu\text{eV}$ (b) $3.78 \times 10^{-28} \text{ J} = 0.236 \text{ neV}$
- 12.17** (a) $1.7 \times 10^{-35} \text{ m}$ (b) $1.1 \times 10^{-32} \text{ m}$ (c) $3.0 \times 10^{-23} \text{ m}$
- 12.18** (a) $6.63 \times 10^{-25} \text{ kg m/s}$ (for both) (b) 1.24 keV (c) 1.51 eV
- 12.19** (a) $6.686 \times 10^{-21} \text{ J} = 4.174 \times 10^{-2} \text{ eV}$ (b) 0.145 nm
- 12.20** $\lambda = h/p = h/(hv/c) = c/v$
- 12.21** α - particle
- 12.22** $1.675 \times 10^{-27} \text{ kg}$, neutron
- 12.23** 0.028 nm
- 12.24** (a) Use $eV = (mv^2/2)$ i.e., $v = [(2eV/m)]^{1/2}$; $v = 1.33 \times 10^7 \text{ m s}^{-1}$
 (b) If we use the same formula with $V = 10^7 \text{ V}$, we get $v = 1.88 \times 10^9 \text{ m s}^{-1}$. This is clearly wrong, since nothing can move with a speed greater than the speed of light ($c = 3 \times 10^8 \text{ m s}^{-1}$). Actually, the above formula for kinetic energy ($mv^2/2$) is valid only when $(v/c) \ll 1$. At very high speeds when (v/c) is comparable to (though always less than) 1, we come to the relativistic domain where the following formulae are valid:

Relativistic momentum $p = m v$

Total energy $E = m c^2$

Kinetic energy $K = m c^2 - m_0 c^2$,

where the relativistic mass m is given by

$$m = m_0 \left(1 - \frac{v^2}{c^2} \right)^{-1/2}$$

m_0 is called the rest mass of the particle. These relations also imply:

$$E = (p^2 c^2 + m_0^2 c^4)^{1/2}$$

Note that in the relativistic domain when v/c is comparable to 1, K or energy $\geq m_0 c^2$ (rest mass energy). The rest mass energy of electron is about 0.51 MeV. Thus a kinetic energy of 10 MeV, being much greater than electron's rest mass energy, implies relativistic domain. Using relativistic formulas, v (for 10 MeV kinetic energy) = 0.999 c .

12.25 (a) 22.7 cm

(b) No. As explained above, a 20 MeV electron moves at relativistic speed. Consequently, the non-relativistic formula $R = (m_0 v / e B)$ is not valid. The relativistic formula is

$$R = \frac{p}{e B} = \frac{m v}{e B}$$

$$\text{or } R = \frac{m_0 v}{e B \sqrt{1 - v^2 / c^2}}$$

12.26 In this case $E = (V/d)$. Therefore, $(e/m) = (V/2B^2 d^2)$. Thus, if V is doubled, B should be increased to $\sqrt{2}B$.

12.27 (a) The deflection at the other edge is given by

$$\begin{aligned} y &= \frac{1}{2} \frac{e E}{m} \times \frac{L^2}{v^2} = \frac{e E L^2}{4 e V} \\ &= \frac{E L^2}{4 V} \end{aligned}$$

where E is the electric field between the plates and V is the accelerating voltage before the beam enters the plates. With the given rules, $y = 5.0$ mm.

(b) The slope of the parabolic trajectory at the edge is given, using the above relation between y and x (replace L by x):

$$\begin{aligned} \tan \theta &= \left. \frac{dy}{dx} \right|_{x=L} \\ &= \frac{E L}{2 V} = \frac{E L^2}{4 V} / (L/2) \end{aligned}$$

which shows that the tangent to the trajectory at the edge when produced back meets the initial direction of the beam at the mid-point $x = L/2$. The same tangent when produced further meets the screen at $y = Y$. Therefore, the deflection at the screen a distance D away from the far edge of the plates is given by:

$$Y = \left(D + \frac{L}{2} \right) \tan \theta$$

$$= \frac{EL}{2V} \left(D + \frac{L}{2} \right)$$

with the given values, $Y = 6.5 \text{ cm}$

$$12.28 \quad Y = \frac{eEL}{2eV} \left(D + \frac{L}{2} \right)$$

$$= \frac{eEL}{mv^2} \left(D + \frac{L}{2} \right)$$

Using the 'no deflection' condition

$$v = \left| \frac{E}{B} \right|, \text{ this gives}$$

$$\frac{e}{m} = \frac{YE}{B^2 L [D + (L/2)]}$$

Using the given data,

$$\frac{e}{m} = 1.7 \times 10^{11} \text{ C kg}^{-1}$$

12.29 We have $eV = (mv^2/2)$ and $R = (mv/eB)$ which gives $(e/m) = (2V/R^2 B^2)$; using the given data $(e/m) = 1.73 \times 10^{11} \text{ C kg}^{-1}$.

$$12.30 \quad (a) \text{ Use } r^2 = \frac{9}{2} \frac{\eta v}{(\rho - \sigma)g}$$

to get $r = 7.26 \times 10^{-7} \text{ m}$ (ρ = density of drop, σ = density of air)

(b) Use $qE = \frac{4\pi}{3} r^3 (\rho - \sigma)g$ to get $q = 8.05 \times 10^{-19} \text{ C}$. From the known value of electronic charge $e = 1.6 \times 10^{-19} \text{ C}$, it is clear that the drop carries 5 excess electrons.

12.31 (a) The radioactive source ionises air through which the drop falls. The drop, therefore, picks up additional charges by acquiring excess electrons from the environment. It may also lose some of its electrons.

$$(b) \quad q_1 E = \frac{4\pi}{3} r^3 (\rho - \sigma)g - 6\pi\eta r v_1$$

$$= 6\pi\eta r (v_0 - v_1)$$

Charge quantisation shows in the discrete values of $(v_0 - v)$ with $v = v_1, v_2, \dots$ i.e., $(v_0 - v)$ are found to be integral multiples of a base value. (Note: If the net motion is upwards, v is negative.)

12.32 Use the relations

$$mg = 6\pi\eta r v_1, \quad qE = 6\pi\eta r v_2$$

where v_1 and v_2 are the vertical and horizontal components of velocity of the drop. Therefore, if θ is the angle with the vertical,

$$\tan \theta = v_2/v_1 = \frac{qE}{mg}$$

$$\text{or } q = \frac{4\pi}{3} \frac{r^3 \rho g}{E} \tan \theta$$

Using the given data, $q = 4.85 \times 10^{-19} \text{ C}$. Clearly, the drop has three excess electrons.

12.33 (a) 27.6 keV

(b) of the order of 30 kV

12.34 Use $\lambda = (hc/E)$ with $E = 5.1 \times 1.602 \times 10^{-10} \text{ J}$ to get $\lambda = 2.43 \times 10^{-16} \text{ m}$.

12.35 (a) For $\lambda = 500 \text{ m}$, $E = (hc/\lambda) = 3.98 \times 10^{-28} \text{ J}$. Number of photons emitted per second

$$\begin{aligned} & \frac{10^4 \text{ Js}^{-1}}{3.98 \times 10^{-28} \text{ J}} \\ & = 3 \times 10^{31} \text{ s}^{-1} \end{aligned}$$

We see that the energy of a radiophoton is exceedingly small, and the number of photons emitted per second in a radio beam is enormously large. There is, therefore, negligible error involved in ignoring the existence of a minimum quantum of energy (photon) and treating the total energy of a radio wave as continuous.

(b) For $\nu = 6 \times 10^{14} \text{ Hz}$, $E = 4 \times 10^{-19} \text{ J}$. Photon flux corresponding to minimum intensity

$$\begin{aligned} & \frac{10^{-10} \text{ W m}^{-2}}{4 \times 10^{-19} \text{ J}} \\ & = 2.5 \times 10^8 \text{ m}^{-2} \text{ s}^{-1} \end{aligned}$$

Number of photons entering the pupil per second $= 2.5 \times 10^8 \times 0.4 \times 10^{-4} \text{ s}^{-1} = 10^4 \text{ s}^{-1}$. Though this number is not as large as in (a) above, it is large enough for us never to 'sense' or 'count' individual photons by our eye.

12.36 $\phi_0 = h\nu - eV_0 = 6.7 \times 10^{-19} \text{ J} = 4.2 \text{ eV}$; $\nu_0 = \frac{\phi_0}{h} = 1.0 \times 10^{15} \text{ Hz}$; $\lambda = 6328 \text{ \AA}$

corresponds to $\nu = 4.7 \times 10^{14} \text{ Hz} < \nu_0$. The photo-cell will not respond howsoever high be the intensity of laser light.

12.37 Use $eV_0 = h\nu - \phi_0$ for both sources. From the data on the first source, $\phi_0 = 1.40 \text{ eV}$. Use this value to obtain for the second source $V_0 = 1.50 \text{ V}$.

12.38 Obtain V_0 versus ν plot. The slope of the plot is (h/e) and its intercept on the ν -axis is ν_0 . The first four points lie nearly on a straight line which intercepts the ν -axis at $\nu_0 = 5.0 \times 10^{14} \text{ Hz}$. The fifth point corresponds to $\nu < \nu_0$; there is no photoelectric emission and therefore no stopping voltage is required to stop the current. Slope of the plot is found to be $4.15 \times 10^{-15} \text{ V s}$. Using $e = 1.6 \times 10^{-19} \text{ C}$, $h = 6.64 \times 10^{-34} \text{ J s}$ (standard value $h = 6.626 \times 10^{-34} \text{ J s}$), $\phi_0 = h\nu_0 = 2.11 \text{ V}$.

12.39 It is found that the given incident frequency ν is greater than $\nu_0(\text{Na})$, and $\nu_0(\text{K})$; but less than $\nu_0(\text{Mo})$, and $\nu_0(\text{Ni})$. Therefore, Mo and Ni will not give photoelectric emission. If the laser is brought closer, intensity of radiation increases, but this does not affect the result regarding Mo and Ni. However, photoelectric current from Na and K will increase in proportion to intensity.

12.40 Assume one conduction electron per atom. Effective atomic area $\sim 10^{-20} \text{ m}^2$. Number of electrons in 5 layers

$$\begin{aligned} & \frac{5 \times 2 \times 10^{-4} \text{ m}^2}{10^{-20} \text{ m}^2} \\ & = 10^{17} \end{aligned}$$

Incident power

$$= 10^{-5} \text{ W m}^{-2} \times 2 \times 10^{-4} \text{ m}^2$$

$$= 2 \times 10^{-9} \text{ W}$$

In the wave picture, incident power is uniformly absorbed by all the electrons continuously. Consequently, energy absorbed per second per electron

$$= \frac{2 \times 10^{-9}}{10^{17}} = 2 \times 10^{-26} \text{ W}$$

Time required for photoelectric emission

$$= \frac{2 \times 1.6 \times 10^{-19} \text{ J}}{2 \times 10^{-26} \text{ W}} = 1.6 \times 10^7 \text{ s}$$

which is about 0.5 year.

Implication: Experimentally, photoelectric emission is observed nearly instantaneously ($\sim 10^{-9}$ s); Thus, the wave picture is in gross disagreement with experiment. In the photon-picture, energy of the radiation is not continuously shared by all the electrons in the top layers. Rather, energy comes in discontinuous 'quanta', and absorption of energy does not take place gradually. A photon is either not absorbed, or absorbed by an electron nearly instantly.

- 12.41** (a) The conditions of energy and momentum conservations are:

$$m_0 c^2 + h\nu = \sqrt{m_0^2 c^4 + p^2 c^2}$$

$$0 + \frac{h\nu}{c} = p$$

These conditions imply: $2m_0 c^3 p = 0$, which is impossible. In a frame where the initial electron is moving with uniform velocity, the same conclusion must hold because if a process is forbidden in one inertial frame, it is also forbidden in another inertial frame.

- (b) We have shown in (a) that

$$e^- + \gamma \rightarrow e^-$$

(forbidden). However, for an electron in a lattice, the momentum of the incident photon can be shared by both the electron and the lattice, while the lattice due to its very large mass (compared to the mass of the electron) does not share the energy of the incident photon. This situation is like a ball rebounding from a wall, where the wall shares momentum but not energy. (Chapter 6, Class XI). In short, while $e^- + \gamma \rightarrow e^-$ is forbidden,

$e^- + \gamma + \text{lattice} \rightarrow e^- + \text{lattice}$, is not forbidden.

- 12.42** Use $h\nu - B = E$; B = binding energy of the level from which electron is emitted after absorption of photon. E is the energy of the photoelectron emitted. For a given ν , E is discrete because B is discrete. Now $B + E = 347 \text{ keV}$ (K shell), 352 keV (L shell), 357.5 keV (M shell). The values of $B + E$ are equal within about 3%. Taking $B + E \approx 350 \text{ keV}$, the wavelength for γ -ray is estimated to be $\lambda = hc/(B + E) = 3.5 \times 10^{-12} \text{ m}$.

- 12.43** Use $R = \alpha E$ to obtain

$$E_1 = \frac{R_1}{\alpha} = \frac{1.40 \text{ cm}}{1 \text{ cm keV}^{-1}}$$

$$= 1.40 \text{ keV}$$

Similarly, $E_2 = 2.02 \text{ keV}$

Incident photon energy $= (hc/\lambda) = 2.53 \text{ keV}$

Use $h\nu - B = E$ to get

$$B_1 = 1.13 \text{ keV}, B_2 = 0.51 \text{ keV}$$

- 12.44** For $\lambda = 1 \text{ \AA}$, electron's energy $= 150 \text{ eV}$; photon's energy $= 12.4 \text{ keV}$. Thus, for the same wavelength, a photon has much greater energy than an electron.

- 12.45** (a) $\lambda = \frac{h}{p} = \frac{h}{\sqrt{2mK}}$ Thus, for same K , λ decreases with m as $(1/\sqrt{m})$. Now

$(m_n/m_e) = 1838.6$; therefore for the same energy, (150 eV) as in Exercise 12.44, wavelength of neutron $= (\sqrt{1838.6}) \times 10^{-10} \text{ m} = 2.33 \times 10^{-12} \text{ m}$. The interatomic spacing is about a hundred times greater. A neutron beam of 150 eV energy is therefore not suitable for diffraction experiments.

- (b) $\lambda = 1.45 \times 10^{-10} \text{ m}$ [Use $\lambda = (h/\sqrt{3mkT})$] which is comparable to interatomic spacing in a crystal. Clearly, from (a) and (b) above, thermal neutrons are a suitable probe for diffraction experiments; so a high energy neutron beam should be first thermalised before using it for diffraction.

12.46 $\lambda = 5.5 \times 10^{-12} \text{ m}$

λ (yellow light) $= 5.9 \times 10^{-7} \text{ m}$

RP is inversely proportional to wavelength. Thus, resolving power (RP) of an electron microscope is about 10^5 times that of an optical microscope. In practice, differences in other (geometrical) factors can change this comparison somewhat.

- 12.47** Use

$$p = \frac{h}{\lambda} = \frac{6.63 \times 10^{-34} \text{ Js}}{10^{-15} \text{ m}} \\ = 6.63 \times 10^{-19} \text{ kg m s}^{-1}$$

Use the relativistic formula for energy:

$$E^2 = c^2 p^2 + m_0^2 c^4 \\ = 9 \times (6.63)^2 \times 10^{-22} + (0.511 \times 1.6)^2 \times 10^{-26} \\ \approx 9 \times (6.63)^2 \times 10^{-22},$$

the second term (rest mass energy) being negligible.

Therefore, $E = 1.989 \times 10^{-10} \text{ J} = 1.24 \text{ BeV}$. Thus, electron energies from the accelerator must have been of the order of a few BeV.

- 12.48** For $\lambda = 10^{-14} \text{ m}$, $E = 124 \text{ MeV}$. This is much too large compared to the binding energy that Coulomb force can provide within the nucleus. Therefore, electrons localised within a nucleus are far too energetic to stay bound within. This is why electrons do not reside in a nucleus.

- 12.49** Use

$$\lambda = \frac{h}{\sqrt{3mkT}} ;$$

$$m_{\text{He}} = \frac{4 \times 10^{-3}}{6 \times 10^{23}} \text{ kg}$$

This gives $\lambda = 0.73 \times 10^{-10} \text{ m}$. Mean separation

$$r = \left(\frac{V}{N} \right)^{1/3} = \left(\frac{kT}{p} \right)^{1/3}$$

For $T = 300 \text{ K}$, $p = 1.01 \times 10^5 \text{ Pa}$, $r = 3.4 \times 10^{-9} \text{ m}$. We find $r \gg \lambda$.

12.50 Using the same formula as in Exercise 12.49, $\lambda = 6.2 \times 10^{-9} \text{ m}$ which is much greater than the given inter-electron separation.

12.51 (a) Quarks are thought to be confined within a proton or neutron by forces which grow stronger if one tries to pull them apart. It, therefore, seems that though fractional charges may exist in nature, observable charges are still integral multiples of e .

(b) Electric fields needed in the experiment will be impractically high.

(c) Stokes' formula is valid for motion through a homogeneous continuous medium. The size of the drop should be much larger than the intermolecular separation in the medium for this assumption to be valid; otherwise the drop 'sees' inhomogeneities in the medium (concentrated mass density in molecules, and holes in between molecules)

(d) Both the basic relations $eV = \frac{1}{2}mv^2$ or $eE = ma$ and $eBv = \frac{mv^2}{r}$,

for electric and magnetic fields, respectively, show that the dynamics of electrons is determined not by e , and m separately but by the combination e/m .

(e) At low pressures, ions have a chance to reach their respective electrodes and constitute a current. At ordinary pressures, ions have no chance to do so because of collisions with gas molecules and recombination.

(f) Work function merely indicates the minimum energy required for the electron in the highest level of the conduction band to get out of the metal. Not all electrons in the metal belong to this level. They occupy a continuous band of levels. Consequently, for the same incident radiation, electrons knocked off from different levels come out with different energies.

(g) The absolute value of energy E (but not momentum p) of any particle is arbitrary to within an additive constant. Hence, while λ is physically significant, absolute value of v of a matter wave of an electron has no direct physical meaning. The phase speed $v\lambda$ is likewise not physically significant. The group speed given by

$$\begin{aligned} \frac{dv}{d(1/\lambda)} &= \frac{dE}{dp} \\ &= \frac{d}{dp} \left(\frac{p^2}{2m} \right) = \frac{p}{m} \end{aligned}$$

is physically meaningful.

Chapter 13

13.1 $2.3 \times 10^{-14} \text{ m}$

13.2 $1.8 \times 10^{-14} \text{ m}$

13.3 $5.6 \times 10^{14} \text{ Hz}$

13.5 $13.6 \text{ eV}; -27.2 \text{ eV}$

13.6 $2.12 \times 10^{-10} \text{ m}; 4.77 \times 10^{-10} \text{ m}$

13.7 5×10^{45}

13.8 621 nm

13.9 We have

$$N_x = N_0 e^{-(E_x - E_0)/kT}$$

$$N_{15}/N_{13} = e^{-(E_{15} - E_{13})/kT}$$

$$= \exp [-(2 \times 1.3 \text{ eV})/kT]$$

$$= \exp \left[- \left(\frac{2.6 \times 1.6 \times 10^{-19}}{1.38 \times 10^{-23} \times 2500} \right) \right]$$

$$= \exp[-12]$$

$$= 1/e^{12}$$

$$= 6.15 \times 10^{-6}$$

13.10 $N_x = N_0 e^{-(E_x - E_0)/kT}$

$$\frac{N_x}{N_0} = e^{-(E_x - E_0)/kT}$$

$$\ln\left(\frac{N_x}{N_0}\right) = -(E_x - E_0)/kT$$

$$\text{or } (-T) = \frac{E_x - E_0}{k \ln\left(\frac{N_x}{N_0}\right)}$$

$$(E_x - E_0) = 2.2 \text{ eV}$$

$$\frac{N_x}{N_0} = 1.10$$

$$(-T) = \frac{2.2 \text{ eV}}{8.62 \times 10^{-5} (\text{eV/K}) \times \ln(1.10)}$$

$$= 2.67 \times 10^5 \text{ K}$$

13.11 3013 m

13.12 24.84 keV

13.14 (a) 69.5 keV ; (b) 17.9 pm

- 13.15** (a) No different from
 (b) Thomson's model; Rutherford's model
 (c) Rutherford's model
 (d) Thomson's model; Rutherford's model
 (e) Both the models

- 13.16** (a) About the same.
 (b) Much less.
 (c) It suggests that the scattering is predominantly due to a single collision, because the chance of a single collision increases linearly with the number of target atoms, and hence linearly with thickness.
 (d) In Thomson's model, a single collision causes very little deflection. The observed average scattering angle can be explained only by considering multiple scattering. So it is wrong to ignore multiple scattering in Thomson's model. In Rutherford's model, most of the scattering comes through a single collision and multiple scattering effects can be ignored as a first approximation.

- 13.17** Angular momentum conservation gives $m vb = m v' s$ (Note at the point of minimum distance s , velocity is normal to the radius vector from the nucleus to the point). Energy conservation gives

$$\frac{1}{2} m v^2 = \frac{1}{2} m v'^2 + \frac{Z Z' e^2}{4 \pi \epsilon_0 s}$$

Eliminate v' to get

$$s^2 = \frac{Z Z' e^2}{2 \pi \epsilon_0 m v^2} + b^2$$

For

$$b = 0, s = r_0 = \frac{Z Z' e^2}{4 \pi \epsilon_0 (m v^2 / 2)}$$

- 13.18** (a) $b = 0$ implies $\cot(\theta/2) = 0$ i.e., $\theta/2 = 90^\circ$ or $\theta = 180^\circ$, as expected physically.
 (b) For a given b , increase in energy implies increase in $\cot(\theta/2)$, and hence decrease in scattering angle.
 (c) $b = 1.1 \times 10^{-14}$ m
 (d) Charge of the nucleus is what provides the field due to which scattering takes place. If $Z = 0$, $\theta = 0$ from the formula, as expected. Mass of the nucleus does not enter since recoil of the nucleus is being ignored. If recoil is included, a small correction term to the formula will include mass of the nucleus.
 (e) For a given energy, decrease in b implies decrease in $\cot(\theta/2)$ and hence increase in scattering angle.

- 13.19** The first orbit Bohr's model has a radius a_0 given by $a_0 = \frac{4 \pi \epsilon_0 (\hbar / 2 \pi)^2}{m_e e^2}$. If we consider the atom bound by the gravitational force ($G m_p m_e / r^2$), we should replace $(e^2 / 4 \pi \epsilon_0)$ by $G m_p m_e$. That is, the radius of the first Bohr orbit is

given by $a_0^G = \frac{(\hbar / 2 \pi)^2}{G m_p m_e} \cong 1.2 \times 10^{29}$ m. This is much greater than the estimated size of the whole universe!

- 13.20** (i) The energy required to excite the electron from the ground state ($n = 1$) to the $n = 3$ state is $E_3 - E_1$

$$\begin{aligned} &= \frac{m Z^2 e^4}{32 \pi^2 \epsilon_0^2 (\hbar / 2 \pi)^2} \left(\frac{1}{1^2} - \frac{1}{3^2} \right) \\ &= 48.4 \text{ eV} \end{aligned}$$

where $Z = 2$.

(ii) Ionisation energy is given by

$$E_{\infty} - E_1$$

$$= \frac{m Z^2 e^4}{32 \pi^2 \epsilon_0^2 (h/2\pi)^2}; \text{ with } Z = 3$$

$$= 122 \text{ eV}$$

Ionisation potential is 122 V.

$$\text{13.21 } v = \frac{m e^4}{(4\pi)^3 \epsilon_0^2 (h/2\pi)^3} \left[\frac{1}{(n-1)^2} - \frac{1}{n^2} \right]$$

$$= \frac{m e^4 (2n-1)}{(4\pi)^3 \epsilon_0^2 (h/2\pi)^3 n^2 (n-1)^2}$$

$$\text{For large } n, v \cong \frac{m e^4}{32 \pi^3 \epsilon_0^2 (h/2\pi)^3 n^3}$$

Orbital frequency $v_c = (v/2\pi r)$. In Bohr model $v = \frac{n(h/2\pi)}{m r}$, and

$$r = \frac{4\pi \epsilon_0 (h/2\pi)^2}{m e^2} n^2. \text{ This gives}$$

$$v_c = \frac{n(h/2\pi)}{2\pi m r^2}$$

$$= \frac{m e^4}{32 \pi^3 \epsilon_0^2 (h/2\pi)^3 n^3}$$

which is same as v for large n .

13.22 (a) The quantity $\left(\frac{e^2}{4\pi \epsilon_0 m c^2} \right)$ has the dimensions of length. Its value is

$2.82 \times 10^{-15} \text{ m}$ – much smaller than the typical atomic size.

(b) The quantity $\frac{4\pi \epsilon_0 (h/2\pi)^2}{m e^2}$ has the dimensions of length. Its value is

$0.53 \times 10^{-10} \text{ m}$ – of the order of atomic sizes. (Note that the dimensional arguments cannot, of course, tell us that we should use 4π and $h/2\pi$ in place of h to arrive at the right size.)

13.23 Ground state energy of a hydrogen atom = -13.6 eV .

Energy of each electron in helium ground state = $-Z_{\text{eff}}^2 \times 13.6 \text{ eV}$.

Ground state energy of a helium atom = $-Z_{\text{eff}}^2 \times 27.2 \text{ eV}$.

Energy of He^+ ground state = $-4 \times 13.6 \text{ eV}$

$$= -54.4 \text{ eV}$$

(Note in He^+ ground state, $Z = 2$, there is no screening because there is only one electron)

$$\begin{aligned}\text{Ionisation potential} &= [-54.4 + Z_{\text{eff}}^2 \times 27.2] \\ &= 24.46 \text{ V (experimental value)}\end{aligned}$$

This gives $Z_{\text{eff}} = 1.70$.

- 13.24** In an ordinary atom, as a first approximation, the motion of the nucleus can be ignored. In a positronium atom, a positron replaces proton of hydrogen atom. The electron and positron masses are equal and, therefore, the motion of the positron cannot be ignored. One must consider the motion of both electron and positron about their center of mass. A detailed analysis (beyond the scope of this book) shows that formulae of Bohr model apply to positronium atom provided that we replace m_e by what is known as the reduced mass of the electron. For positronium, the reduced mass is $m_e/2$. The wavelength of the radiation emitted in the $n = 2$ to $n = 1$ transition is double than that of the corresponding radiation emitted for a similar transition in hydrogen atom which has a wavelength of 1217 \AA and hence is equal to $2 \times 1217 = 2434 \text{ \AA}$. This radiation is in the ultraviolet part of the em spectrum.

13.25
$$r_n = \frac{4\pi\epsilon_0\hbar^2}{Ze^2m} n^2 \text{ i.e., } r_n \propto \frac{n^2}{Z}.$$

For hydrogen, $Z=1$, for Be^{++} , $Z=4$. The ($n=2$) state of Be^{++} has the same radius as ($n=1$) state of hydrogen. Now $E_n \propto Z^2/n^2$. Therefore $[(E_2(\text{Be}^{++})/E_1(\text{H}))] = 4$.

- 13.26** $E_n \propto Z^2/n^2$ for Li^{++} , $Z=3$. Therefore, ($n=3$ state of Li^{++} has the same energy as ($n=1$) state of hydrogen. Now $r_n \propto (n^2/Z)$. Therefore, $r_3(\text{Li}^{++}) = 3r_1(\text{H})$.

13.27 In Bohr's model, $mvr = n\hbar$ and $\frac{mv^2}{r} = \frac{Ze^2}{4\pi\epsilon_0r^2}$

which give

$$T = \frac{1}{2}mv^2 = \frac{Ze^2}{8\pi\epsilon_0r};$$

$$r = \frac{4\pi\epsilon_0\hbar^2}{Ze^2m} n^2$$

These relations have nothing to do with the choice of the zero of potential energy. Now, choosing the zero of potential energy at infinity we have $V = -(Ze^2/4\pi\epsilon_0r)$ which gives $V = 2T$ and $E = T + V = -T$

- (a) The quoted value of $E = -3.4 \text{ eV}$ is based on the customary choice of zero of potential energy at infinity. Using $E = -T$, the kinetic energy of the electron in this state is $+3.4 \text{ eV}$.
- (b) Using $V = -2T$, potential energy of the electron is $= -6.8 \text{ eV}$
- (c) If the zero of potential energy is chosen differently, kinetic energy does not change. Its value is $+3.4 \text{ eV}$ independent of the choice of the zero of potential energy. The potential energy, and the total energy of the state, however, would alter if a different zero of the potential energy is chosen.

13.28
$$v = \frac{n\hbar}{mr}, r = \frac{4\pi\epsilon_0\hbar^2n^2}{Ze^2m}$$

That is
$$v = \frac{Ze^2}{4\pi\epsilon_0n\hbar}$$

For $n = 3$, $Z = 2$, $v = 1.46 \times 10^6 \text{ m s}^{-1}$

Thus, $(v/c \approx 0.005)$; non-relativistic approximation is valid because $(v/c) \ll 1$.

13.29 Angular momenta associated with planetary motion are incomparably large relative to \hbar . For example, angular momentum of the earth in its orbital motion is of the order of $10^{70}\hbar$. In terms of the Bohr's quantisation postulate, this corresponds to a very large value of n (of the order of 10^{70}). For such large values of n , the differences in the successive energies and angular momenta of the quantised levels of the Bohr model are so small compared to the energies and angular momenta, respectively, for the levels that one can, for all practical purposes, consider the levels continuous.

13.30 All that is needed is to replace m_e by m_μ in the formulas of the Bohr model. We note that keeping other factors fixed, $r \propto (1/m)$ and $E \propto m$. Therefore,

$$\begin{aligned} r_\mu &= \frac{r_e m_e}{m_\mu} = \frac{0.53 \times 10^{-13}}{207} \\ &= 2.56 \times 10^{-13} \text{ m} \\ E_\mu &= \frac{E_e m_\mu}{m_e} = -(13.6 \times 207) \text{ eV} \\ &\approx -2.8 \text{ keV} \end{aligned}$$

Chapter 14

14.1 20.18 u

14.2 104.7 MeV

14.3 1.6×10^{25} MeV: 1 gram-mole of a substance contains 6×10^{23} atoms.

14.4 $1 \text{ u} = 1.660565 \times 10^{-27} \text{ kg}$
 $1 \text{ u} \times c^2 \approx 931.5 \text{ MeV}$

Using the formula for binding energy given in Section 14.4, we get

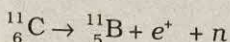
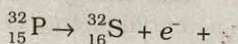
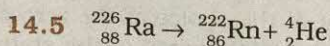
$$\text{B.E. } ({}^{56}_{26}\text{Fe}) = 492.26 \text{ MeV}$$

$$\text{B.E. per nucleon} = 8.79 \text{ MeV}$$

$$\text{B.E. } ({}^{209}_{83}\text{Bi}) = 1640.30 \text{ MeV}$$

$$\text{B.E. per nucleon} = 7.85 \text{ MeV}$$

${}^{56}_{26}\text{Fe}$ has greater binding energy per nucleon.



14.6 4 T years

14.7 7.1 mg

14.8 ${}^{238}_{92}\text{U} \rightarrow {}^{234}_{90}\text{Th} + {}^4_2\text{He} + Q$, where Q represents the kinetic energy released.

From Einstein's mass-energy equivalence

$$Q = [m_N(^{238}_{92}\text{U}) - m_N(^{234}_{90}\text{Th}) - m_N(^4_2\text{He})]c^2$$

Add and subtract $92 m_e$ in the bracket to write Q in terms of atomic masses (ignoring mass defect due to electronic binding energies):

$$Q = [m(^{238}_{92}\text{U}) - m(^{234}_{90}\text{Th}) - m(^4_2\text{He})]c^2$$

Using the given data and $c^2 = 931.5 \text{ MeV/u}$, $Q = 4.27 \text{ MeV}$. This total kinetic energy is shared between the $^{234}_{90}\text{Th}$ recoil nucleus and the α -particle. In the rest frame of $^{238}_{92}\text{U}$, $^{234}_{90}\text{Th}$ and the α -particle will have equal and opposite momenta. Consequently, since $^{234}_{90}\text{Th}$ is very much massive as compared to α -particle, the kinetic energy of $^{234}_{90}\text{Th}$ nucleus is much less than that of α -particle. Thus, most of the kinetic energy released is carried away by the α -particle.

- 14.9** $^{11}_6\text{C} \rightarrow ^{11}_5\text{B} + e^+ + n + Q$, where Q is the kinetic energy released in the decay process. The Q of this decay process is given by

$Q = [m_N(^{11}_6\text{C}) - m_N(^{11}_5\text{B}) - m_e]c^2$; the masses used in this equation are nuclear masses.

Now, if we express the Q value in terms of atomic masses we have to subtract $6m_e$ from the atomic mass carbon and $5m_e$ from that of boron atomic to get the corresponding nuclear masses. Therefore, we have

$$\begin{aligned} Q &= [m(^{11}\text{C}) - 6m_e - m(^{11}\text{B}) + 5m_e - m_e]c^2 \\ &= [m(^{11}\text{C}) - m(^{11}\text{B}) - 2m_e]c^2 \end{aligned}$$

Now, using the atomic mass data we get

$$Q = 0.961 \text{ MeV.}$$

- 14.10** $^{23}_{10}\text{Ne} \rightarrow ^{23}_{11}\text{Na} + e^- + \bar{\nu} + Q$

$$Q = [m_N(^{23}_{10}\text{Ne}) - m_N(^{23}_{11}\text{Na}) - m_e]c^2$$

where the neutrino mass has been neglected. Thus,

$$\begin{aligned} Q &= [m(^{23}_{10}\text{Ne}) - 10m_e - m(^{23}_{11}\text{Na}) + 11m_e - m_e]c^2 \\ &= [m(^{23}_{10}\text{Ne}) - m(^{23}_{11}\text{Na})]c^2 \\ &= 4.374 \text{ MeV} \end{aligned}$$

This is the maximum energy of the β^- emitted.

- 14.11** (i) $Q = [m_N(^1_1\text{H}) + m_N(^3_1\text{H}) - 2m_N(^2_1\text{H})]c^2$
 $= [m(^1_1\text{H}) + m(^3_1\text{H}) - 2m(^2_1\text{H})]c^2$
 $= -4.03 \text{ MeV}$

(ii) $Q = [2m_N(^{12}_6\text{C}) - m_N(^{20}_{10}\text{Ne}) - m_N(^4_2\text{He})]c^2$

Reaction (i) is endothermic, while reaction (ii) is exothermic.

- 14.12** The disintegration energy in the fission of $^{98}_{42}\text{Mo}$ is given by

$$\begin{aligned} Q &= [m_N(^{98}_{42}\text{Mo}) + m_n - 2m_N(^{49}_{21}\text{Sc})]c^2 \\ &= [m(^{98}_{42}\text{Mo}) + m_n - 2m(^{49}_{21}\text{Sc})]c^2 \\ &= 944.6 \text{ MeV} \end{aligned}$$

14.13 4.5×10^{23} MeV

14.14 180 keV

14.15 4.9×10^4

14.16 ^{25}Mg : 9.303%; ^{26}Mg : 11.71%

14.17 Neutron separation energy S_n of a nucleus ^A_ZX is given by

$$S_n = [m_N(^{A-1}_Z\text{X}) + m_n - m_N(^A_Z\text{X})]c^2$$

From the given data, using $c^2 = 931.5$ MeV/u, we get

$$S_n(^{41}_{20}\text{Ca}) = 8.36 \text{ MeV}$$

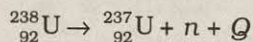
$$S_n(^{27}_{13}\text{Al}) = 13.06 \text{ MeV}$$

14.18 209 d

14.19 (a) For answer see section 14.7.2.

(Measure of the activity of a radioactive source).

(b) The probable decay of $^{238}_{92}\text{U}$ by neutron emission can be written as:



$$Q = [m_N(^{238}_{92}\text{U}) - m_N(^{237}_{92}\text{U}) - m_n]c^2$$

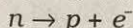
$$= [m(^{238}_{92}\text{U}) - m(^{237}_{92}\text{U}) - m_n]c^2$$

$$= -0.00659 \text{ u}$$

$$= -6.14 \text{ MeV}$$

As the Q value is negative, spontaneous decay of $^{238}_{92}\text{U}$ by neutron emission is not possible.

14.20 (a) Consider the decay of a free neutron at rest:



(i) Since there are only two particles in the final state, conservation of momentum requires that the momenta of electron and proton must be equal and opposite. If the momentum of the electron is p_e , then the momentum of proton should be $-p_e$. Next, by total energy conservation we have,

$$\sqrt{c^2 p_e^2 + m_e^2 c^4} + \sqrt{c^2 p_p^2 + m_p^2 c^4} = m_n c^2$$

(ii) The kinetic energy of the electron is given by

$$T_e = \frac{p_e^2}{2m_e}$$

and that of the proton is

$$T_p = \frac{p_p^2}{2m_p}$$

Thus, there is a definite momentum p_e given by Eq.(ii) in terms of the masses of the particles involved and hence the kinetic energies of the electron and proton are also fixed. Therefore, in the above decay process the electrons cannot have a continuous distribution of energies. The presence of an additional particle in the decay process

allows this possibility. Three particles – the electron, proton and the third particle – will then share the available energy. The electron energy is then no longer fixed. This simple reasoning was among the several arguments that led Pauli to explain the observed continuous distribution of electron energy in β -decay by postulating the existence of a new particle till then unobserved. We now know that the correct equation for β -decay is:

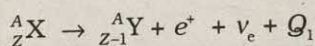
$$n \rightarrow p + e^- + \bar{\nu}_e$$

where the particle denoted by $\bar{\nu}_e$ is called the (electron) antineutrino.

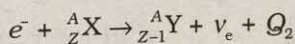
It is a neutral particle of negligibly small rest mass and intrinsic spin $\frac{1}{2}$.

- (b) A free neutron has rest mass greater than that of a proton plus electron. Thus, β^- -decay is energetically allowed, but β^+ -decay of a free proton is not allowed: In a nucleus, individual neutrons and protons are not free. Thus the β^+ -decay of a proton ($p \rightarrow n + e^+ + \nu_e$) is possible when the proton is bound in a nucleus. In a stable nucleus with Z protons and $(A - Z)$ neutrons, the two reciprocal processes (neutron decay and proton decay) are in dynamical equilibrium.

14.21 Consider the two competing processes:



(positron emission)



(electron capture)

$$Q_1 = [m_N({}_Z^AX) - m_N({}_{Z-1}^AY) - m_e]c^2$$

$$= [m({}_Z^AX) - Z m_e - m({}_{Z-1}^AY) - (Z - 1)m_e - m_e]c^2$$

$$= [m({}_Z^AX) - m({}_{Z-1}^AY) - 2m_e]c^2$$

$$Q_2 = [m_N({}_Z^AX) + m_e - m_N({}_{Z-1}^AY)]c^2$$

$$= [m({}_Z^AX) - Z m_e + m_e - m_N({}_{Z-1}^AY) + (Z - 1)m_e]c^2$$

$$= [m({}_Z^AX) - m_N({}_{Z-1}^AY)]c^2$$

This means that $Q_1 > 0$ implies $Q_2 > 0$ but $Q_2 > 0$ does not necessarily mean $Q_1 > 0$.

Hence the result.

14.22 $N = N_0 e^{-\lambda t}$. The activity is proportional to the number of radioactive atoms, so

$$e^{-\lambda t} = (9/15) \text{ i.e., } t = (1/\lambda) \ln(5/3)$$

λ is related to half-life by $\lambda = 0.693/T_{1/2}$.

Thus,

$$t = (T_{1/2}/0.693) \times \ln(5/3)$$

$$= 4224 \text{ y}$$

14.23 (a) For the decay process ${}_{88}^{223}\text{Ra} \rightarrow {}_{82}^{209}\text{Pb} + {}_6^{14}\text{C} + Q$

$$Q = [m_N({}_{88}^{223}\text{Ra}) - m_N({}_{82}^{209}\text{Pb}) - m_N({}_6^{14}\text{C})]c^2$$

$$= [m({}^{223}_{88}\text{Ra}) - m({}^{209}_{82}\text{Pb}) - m({}^{14}_6\text{C})]c^2$$

$$= 31.85 \text{ MeV}$$

For the decay process ${}^{223}_{88}\text{Ra} \rightarrow {}^{219}_{86}\text{Rn} + {}^4_2\text{He} + Q$

$$Q = [m_N({}^{223}_{88}\text{Ra}) - m_N({}^{219}_{86}\text{Rn}) - m_N({}^4_2\text{He})]c^2$$

$$= [m({}^{223}_{88}\text{Ra}) - m({}^{219}_{86}\text{Rn}) - m({}^4_2\text{He})]c^2$$

$$= 5.98 \text{ MeV}$$

- (b) The Coulomb barrier height for α -decay is equal to the Coulomb repulsion between the α -particle and the daughter nucleus when they are just touching each other and hence given by

$$U(\alpha) = \frac{2 \times 86 e^2}{r_0(219^{1/3} + 4^{1/3})}$$

Similarly, the Coulomb barrier for ${}^{14}\text{C}$ would be given by

$$U({}^{14}\text{C}) = \frac{6 \times 82 e^2}{r_0(209^{1/3} + 6^{1/3})}$$

$$U({}^{14}\text{C}) / U(\alpha) = \frac{6 \times 82 \times (219^{1/3} + 4^{1/3})}{(209^{1/3} + 6^{1/3}) \times 2 \times 88}$$

$$U({}^{14}\text{C}) = U(\alpha) \times \frac{6 \times 82 \times (219^{1/3} + 4^{1/3})}{(209^{1/3} + 6^{1/3}) \times 2 \times 88}$$

$$= 30 \times \frac{6 \times 82 \times (219^{1/3} + 4^{1/3})}{(209^{1/3} + 6^{1/3}) \times 2 \times 88}$$

$$= 86 \text{ MeV}$$

14.24 Energy generated per gram of ${}^{235}_{92}\text{U} = \frac{6 \times 10^{23} \times 200 \times 1.6 \times 10^{-13}}{235} \text{ J g}^{-1}$

The amount of ${}^{235}_{92}\text{U}$ consumed in 5y = $\frac{5 \times 365 \times 24 \times 6 \times 6 \times 10^{11}}{6 \times 2 \times 1.6 \times 10^{12}} \times 235 \text{ g}$

$$= 1930 \text{ kg}$$

The initial amount of ${}^{235}_{92}\text{U} = 3860 \text{ kg}$

14.25 For the fission ${}^{238}_{92}\text{U} + n \rightarrow {}^{140}_{58}\text{Ce} + {}^{99}_{44}\text{Ru} + Q$

$$Q = [m_N({}^{238}_{92}\text{U}) + m_n - m_N({}^{140}_{58}\text{Ce}) - m_N({}^{99}_{44}\text{Ru})]c^2$$

$$= [m({}^{238}_{92}\text{U}) + m_n - m({}^{140}_{58}\text{Ce}) - m({}^{99}_{44}\text{Ru})]c^2$$

$$= 231.1 \text{ MeV}$$

14.26 (a) For the process ${}^2_1\text{H} + {}^3_1\text{H} \rightarrow {}^4_2\text{He} + n + Q$

$$Q = [m_N({}^2_1\text{H}) + m_N({}^3_1\text{H}) - m_N({}^4_2\text{He}) - m_n]c^2$$

$$= [m({}^2_1\text{H}) + m({}^3_1\text{H}) - m({}^4_2\text{He}) - m_n]c^2$$

$$= 17.59 \text{ MeV}$$

(b) Repulsive potential of two nuclei when they almost touch each other

$$\begin{aligned}
 &= \frac{q^2}{4\pi\epsilon_0 d} \\
 &= \frac{9 \times 10^9 \times (1.6 \times 10^{-19})^2}{2 \times 1.5 \times 10^{-15}} \text{ J} \\
 &= 7.68 \text{ J}
 \end{aligned}$$

14.27 (a) In Sun, 4 hydrogen nuclei combine to form a helium nucleus with a release of ~26 MeV of energy.

$$\begin{aligned}
 \text{The energy released by 1 kg of hydrogen} &= \frac{6 \times 10^{23} \times 26}{4} \times 10^3 \text{ MeV} \\
 &= 39 \times 10^{26} \text{ MeV}
 \end{aligned}$$

(b) Energy released in fusion of one atom of ${}^{235}_{92}\text{U} = 200 \text{ MeV}$

$$\begin{aligned}
 \text{Energy released in fusion of 1 kg of } {}^{235}_{92}\text{U} &= \frac{6 \times 10^{23} \times 200}{235} \times 1000 \text{ MeV} \\
 &= 5.1 \times 10^{26} \text{ MeV}
 \end{aligned}$$

The energy released in fusion of 1 kg of hydrogen is about 8 times that of the energy released in the fusion of 1 kg of uranium.

14.28 $1.8 \times 10^6 \text{ y.}$

14.29 $\nu(\gamma_1) = 2.608 \times 10^{20} \text{ Hz}$

$$\nu(\gamma_2) = 9.950 \times 10^{19} \text{ Hz}$$

$$\nu(\gamma_3) = 1.633 \times 10^{20} \text{ Hz}$$

These frequencies are obtained by dividing energy differences by h . Maximum kinetic energy of β -particle:

$$\begin{aligned}
 K_{\max}(\beta_1^-) &= c^2[m({}^{198}_{79}\text{Au}) - \text{mass of second excited state of } {}^{198}_{80}\text{Hg}] \\
 &= c^2[m({}^{198}_{79}\text{Au}) - \{m({}^{198}_{80}\text{Hg}) + \frac{1.088}{931.5}\}]
 \end{aligned}$$

Use $c^2 = 931.5 \text{ MeV/u}$ to get

$$\begin{aligned}
 K_{\max}(\beta_1^-) &= [931.5\{m({}^{198}_{79}\text{Au}) - m({}^{198}_{80}\text{Hg})\} - 1.088] \text{ MeV} \\
 &= 0.281 \text{ MeV.}
 \end{aligned}$$

Similarly, $K_{\max}(\beta_2^-) = 0.957 \text{ MeV.}$

14.30 (a) A chemical equation is balanced in the sense that the number of atoms of each element is the same on both sides of the equation. A chemical reaction merely alters the original combinations of atoms. In a nuclear reaction, elements may be transmuted. Thus, the number of atoms of each element is not necessarily conserved in a nuclear reaction. However, the number of protons and the number of neutrons are both separately conserved in a nuclear reaction. [Actually, even this is not strictly true in the realm of very high energies – what is strictly conserved

is the total charge and total 'baryon number'. We need not pursue this matter here.]

In the reactions of Exercise 14.11, the number of protons and the number of neutrons are the same on the two sides of the equation.

- (b) We know that the binding energy of a nucleus gives a negative contribution to the mass of the nucleus (mass defect). Now, since proton number and neutron number are conserved in a nuclear reaction, the total rest mass of neutrons and protons is the same on either side of a reaction. But the total binding energy of nuclei on the left side need not be the same as that on the right hand side. The difference in these binding energies appears as energy released or absorbed in a nuclear reaction. Since binding energy contributes to mass, we say that the difference in the total mass of nuclei on the two sides get converted into energy or vice versa. It is in these sense that a nuclear reaction is an example of mass-energy interconversion.
- (c) From the point of view of mass-energy interconversion, a chemical reaction is similar to a nuclear reaction *in principle*. The energy released or absorbed in a chemical reaction can be traced to the difference in chemical(not nuclear) binding energies of atoms and molecules on the two sides of a reaction. Since, strictly speaking, chemical binding energy also gives a negative contribution (mass defect) to the total mass of an atom or molecule, we can equally well say that the difference in the total mass of atoms or molecules, on the two sides of the chemical reaction gets converted into energy or vice versa. However, the mass defects involved in a chemical reaction are almost a million times smaller than those in a nuclear reaction. This is the reason for the general impression, (which is *incorrect*) that mass-energy interconversion does not take place in a chemical reaction.

14.31 Required power from nuclear plant = 10^{10} W

Required electric energy from nuclear plant in one year = 3.156×10^{17} J

Required number of fissions per year = 3.945×10^{28}

Available electric energy per fission = $0.25 \times 200 = 50$ MeV = 8×10^{-12} J

Required numbers of moles of ^{235}U = 6.55×10^4

Required mass of ^{235}U = $6.55 \times 235 \times 10^4$ g = 1.54×10^4 kg.

Chapter 15

15.1 (c)

15.2 (d)

15.3 (c)

15.4 (c)

15.5 (c)

15.6 (b), (c)

15.7 (c)

15.8 $4.45 \times 10^{22} \text{ m}^{-3}$ [Hint: $n_{T_1} = Ce^{-E_g/2kT_1}$; $n_{T_2} = Ce^{-E_g/2kT_2}$]

$$\ln \frac{n_{T_2}}{n_{T_1}} = -\frac{E_g}{2k} \left(\frac{1}{T_1} - \frac{1}{T_2} \right); \quad T_1 = 300 \text{ K}, T_2 = 500 \text{ K}$$

$$n \text{ at } 500 \text{ K} = 2 \times 10^{19} \times 2.224 \times 10^3 \\ = 4.45 \times 10^{22} \text{ m}^{-3}]$$

15.9 $n_h = 2 \times 10^4 \text{ m}^{-3}$; $N_D \approx n_e$

(use $n_e n_h = n_i^2$)

15.10 50 Hz for half-wave, 100 Hz for full-wave

15.11 $v_i = 0.01 \text{ V}$; $I_B = 10 \mu\text{A}$

15.12 2 V

15.13 No ($h\nu$ has to be greater than E_g).

15.14 $n_e \approx 4.95 \times 10^{22}$; $n_h = 4.75 \times 10^9$; n-type since $n_e \gg n_h$

Hint : For charge neutrality $N_D - N_A = n_e - n_h$; $n_e n_h = n_i^2$

Solving these equations, $n_e = \frac{1}{2} \left[(N_D - N_A) + \sqrt{(N_D - N_A)^2 + 4n_i^2} \right]$

15.15 OR gate

15.16 (i) NOT, (ii) AND

15.18 NOT ; A Y

0 1

1 0

15.19 (a) AND (b) OR

Chapter 16

16.1 (b) and (c) which travel to *line of sight* distance only.

16.2 (a)

16.3 (c)

16.4 (d)

16.5 (d)

16.6 2.5×10^4

16.7 $\phi_c = \sin^{-1} \left(\frac{1.51}{1.55} \right)$; $(\theta_0)_{\max} = 20.48^\circ$

16.8 E_g has to be less than $h\nu$ or hc/λ . None of the semiconductors is suitable since E_g for $\lambda (= 1400 \text{ nm})$ is approximately 0.89 eV which is less than the least E_g of any of the given semiconductors.

16.9 $1.25 \times 10^{12} \text{ m}^{-3}$

16.10 87.6 km

16.11 94.7 km

16.12 1.21

BIBLIOGRAPHY

TEXT BOOKS

For additional reading on the topics covered in this book, you may like to consult one or more of the following books. Some of these books, however, are more advanced and contain many more topics than this book.

- 1 **Ordinary Level Physics**, A.F. Abbott, Arnold-Heinemann (1984).
- 2 **Advanced Level Physics**, M. Nelkon and P. Parker, 6th Edition, Arnold-Heinemann (1987).
- 3 **Advanced Physics**, Tom Duncan, John Murray (2000).
- 4 **Fundamentals of Physics**, David Halliday, Robert Resnick and Jearl Walker, John Wiley (1997).
- 5 **University Physics**, H.D. Young, M.W. Zemansky and F.W. Sears, Narosa Pub. House (1982).
- 6 **Problems in Elementary Physics**, B. Bukhovtsova, V. Krivchenkov, G. Myakishev and V. Shalnov, MIR Publishers, (1971).
- 7 **Lectures on Physics** (3 volumes), R. P. Feynman, Addison – Wesley (1965).
- 8 **Berkeley Physics Course** (5 volumes) McGraw Hill (1965).
 - a. Vol. 1 – Mechanics: (Kittel, Knight and Ruderman)
 - b. Vol. 2 – Electricity and Magnetism (E.M. Purcell)
 - c. Vol. 3 – Waves and Oscillations (Frank S. Crawford)
 - d. Vol. 4 – Quantum Physics (Wichmann)
 - e. Vol. 5 – Statistical Physics (F. Reif)
- 9 **Fundamental University Physics**, M. Alonso and E. J. Finn, Addison – Wesley (1967).
- 10 **College Physics**, R.L. Weber, K.V. Manning, M.W. White and G.A. Weygand, Tata McGraw Hill (1977).
- 11 **Physics: Foundations and Frontiers**, G. Gamow and J.M. Cleveland, Tata McGraw Hill (1978).

- 12 **Physics for the Inquiring Mind**, E.M. Rogers, Princeton University Press (1960)
- 13 **PSSC Physics Course**: DC Heath and Co. (1965) Indian Edition, NCERT (1967)
- 14 **Physics Advanced Level**, Jim Breithampt, Stanley Thornes Publishers (2000).
- 15 **Physics**, Patrick Fullick, Heinemann (2000).
- 16 **Conceptual Physics**, Paul G. Hewitt, Addison-Wesley (1998).
- 17 **College Physics**, Raymond A. Serway and Jerry S. Faughn, Harcourt Brace and Co. (1999).
- 18 **University Physics**, Harris Benson, John Wiley (1996).
- 19 **University Physics**, William P. Crummet and Arthur B. Western, Wm. C. Brown (1994).
- 20 **General Physics**, Morton M. Sternheim and Joseph W. Kane, John Wiley (1986).
- 21 **Physics**, Hans C. Ohanian, W.W. Norton (1989).
- 22 **Advanced Physics**, Keith Gibbs, Cambridge University Press (1996).
- 23 **Understanding Basic Mechanics**, F. Reif, John Wiley (1995).
- 24 **College Physics**, Jerry D. Wilson and Anthony, J. Buffa, Prentice-Hall (1997).
- 25 **Senior Physics, Part – I**, I.K. Kikoin and A.K.Kikoin, Mir Publishers (1987).
- 26 **Senior Physics, Part – II**, B. Bekhovtsev, Mir Publishers (1988).
27. **Physics of Semi-Conductor Devices**, S.M. Sze, John Wiley (1981).
28. **Transistor Physics and Circuit Design**, D.C. Sarkar, S. Chand and Co. (1985).
29. **The Art of Electronics**, Paul Horowitz and Winfield Hill, Cambridge University Press (1989).
30. **Principles of Electronic Devices**, William D. Stanley, Prentice Hall (1994).
31. **Electronics Fundamentals and Applications**, J.D. Ryder, Prentice Hall of India (1969).
32. **Electronic Communications**, D. Roddy and J. Collen, Prentice Hall of India (1995).
33. **Communication Systems**, Simon Haykin, Wiley Eastern (1985).
34. **Understanding Optical Fibre Communication**, Alan J. Rogers, Artech House (2001).
35. **Optical Fibre Communications**, G. Keiser, McGraw Hill (2000).

GENERAL BOOKS

For instructive and entertaining general reading on science, you may like to read some of the following books. Remember, however, that many of these books are written at a level far beyond the level of the present book.

- 1 **Mr. Tompkins** in paperback, G. Gamow, Cambridge University Press (1967).
- 2 **The Universe and Dr. Einstein**, C. Barnett, Time Inc. New York (1962).
- 3 **Thirty years that shook physics**, G. Gamow, Double Day, New York (1966).
- 4 **Surely You're joking, Mr. Feynman**, R.P. Feynman, Bantam Books (1986).
- 5 **One, Two, Three... Infinity**, G. Gamow, Viking Inc. (1961).
- 6 **The Meaning of Relativity**, A. Einstein, (Indian Edition) Oxford and IBH Publ. Co. (1965).
- 7 **Atomic theory and the Description of Nature**, Niels Bohr, Cambridge (1934).
- 8 **The Physical Principles of Quantum Theory**, W. Heisenberg, University of Chicago Press (1930).
- 9 **The Physics- Astronomy Frontier**, F. Hoyle and J.V. Narlikar, W.H. Freeman (1980).
- 10 **The Flying Circus of Physics with Answer**, J. Walker, John Wiley and Sons (1977).
- 11 **Physics for Everyone** (series), L.D. Landau and A.I. Kitaigorodski, MIR Publisher (1978).
 - Book 1: Physical Bodies
 - Book 2: Molecules
 - Book 3: Electrons
 - Book 4: Photons and Nuclei.
- 12 **Physics can be Fun**, Y. Perelman, MIR Publishers (1986).
- 13 **Power of Ten**, Philip Morrison and Eames, W.H. Freeman (1985).
- 14 **Physics in your kitchen lab.**, I.K. Kikoin, MIR Publishers (1985).

SCIENCE RELATED VALUES

Curiosity, quest for knowledge, objectivity, honesty and truthfulness, courage to question, systematic reasoning, acceptance after proof/verification, open-mindedness, search for perfection and team spirit are some of the basic values related to science. The processes of science, which help in searching the truth about nature and its phenomena are characterised by these values. Science aims at explaining things and events. Therefore to learn and practise science :

- Be inquisitive about things and events around you.
- Have the courage to question beliefs and practices.
- Ask 'what', 'how' and 'why' and find your answers by critically observing, experimenting, consulting, discussing and reasoning.
- Record honestly your observations and experimental results in your laboratory or outside it.
- Repeat experiments carefully and systematically if required, but do not manipulate your results under any circumstance.
- Be guided by facts, reasons and logic. Do not be biased in one way or the other.
- Aspire to make new discoveries and inventions by sustained and dedicated work.



1249



राष्ट्रीय शैक्षिक अनुसंधान और प्रशिक्षण परिषद्
NATIONAL COUNCIL OF EDUCATIONAL RESEARCH AND TRAINING

ISBN 81-7450-194-0